Seale-Carlisle, T. M., Wetmore, S. A., Flowe, H. D., & Mickes, L. (2019). Designing police lineups to maximize memory performance. *Journal of Experimental Psychology: Applied*, *25*(3), 410-430. https://doi.org/10.1037/xap0000222

Designing police lineups to maximize memory performance

Travis M. Seale-Carlisle[1,3], Stacy A. Wetmore[2], Heather D. Flowe[3], & Laura Mickes[1]

[1]Royal Holloway, University of London, [2]Butler University, [3]University of Birmingham

Author Note

This work was supported in part by the Economic and Social Research Council [ES/L012642/1] to Laura Mickes. The content is solely the responsibility of the authors and does not necessarily reflect the views of the Economic and Social Research Council.

We thank 1) the London Metropolitan Police Officers who assisted with the materials used in these experiments; 2) Stanislav Wronski for programming many of the experiments; 3) John T. Wixted and Shiri Lev-Ari for discussions about this research; and 4) Harold Pashler for recommending that our figures reflect the number of responses per point.

Correspondence concerning this article should be addressed to Travis Seale-Carlisle (T.M.Seale-Carlisle@bham.ac.uk) or Laura Mickes (laura.mickes@rhul.ac.uk).

Abstract

How can lineups be designed to elicit the best achievable memory performance? One step toward that goal is to compare lineup procedures. In a recent comparison of US and UK lineup procedures, discriminability and reliability was better when memory was tested using the US procedure. However, because there are so many differences between the procedures, it is unclear what explains this superior performance. The main goal of the current research is therefore to systematically isolate the differences between the US and UK lineups to determine their effects on discriminability and reliability. In five experiments, we compared (1) presentation format: simultaneous vs. sequential; (2) stimulus format: photos vs. videos; (3) number of views: 1-lap vs. 2-lap vs. choice in both video and photo lineups; and (4) lineup size: 6- versus 9-lineup members. Most of the comparisons did not show appreciable differences, but one comparison did: simultaneous presentation yielded better discriminability than sequential presentation. If the results replicate, then policymakers should recommend using a simultaneous lineup procedure. Moreover, consistent with previous research, identifications made with high confidence were higher in reliability than identifications made with low confidence. Thus, official lineup protocols should require collecting confidence because of the diagnostic value added.

Public Significance Statement

We investigated ways to design police lineups so that memory performance is improved. In a set of five experiments testing different aspects of lineups, including simultaneous vs. sequential presentation, photos vs. videos formats, the number of times the lineup is viewed, and the number of 6 vs. 9 lineup members. Results from over 14,000 participants showed that the key factor that improved memory accuracy is to present lineup members

simultaneously. Confidence is an indicator of accuracy and thus improves the probative value of eyewitness evidence.

Keywords: Eyewitness identification, discriminability, confidence-accuracy, sequential lineup, simultaneous lineup

Designing police lineups to maximize memory performance

When poor eyewitness identification procedural practices are used, innocent suspects are endangered. Minimizing that risk has been the focus of much past research, sometimes at the expense of increasing the chances of exonerating guilty suspects. Instead of managing the trade-offs associated with changing response bias (i.e., changing the likelihood of choosing someone from a lineup), which increases one type of error while decreasing the other type of error, the goal should be to find procedures that decrease both types of errors. In other words, the goal should be to decrease identifications of innocent suspects and to increase identifications of guilty suspects (Clark, 2012). Increasing the ability of eyewitnesses to discriminate innocent from guilty suspects accomplishes that goal (Wixted & Mickes, 2012).

Identification procedures vary from country to country, and even within countries they often vary from jurisdiction to jurisdiction. They vary on a number of dimensions, including different presentation types (e.g., sequential or simultaneous), different formats (live, photo, video, or computer-generated), and different sizes (from 1- to 12-member lineups). If the lineup members are sequentially presented, the stopping rules and number of laps allowed vary. This variability led us to ask what procedural design leads to the best performance? And relatedly, what is it about that procedure that accounts for its better performance?

One step toward designing an optimal identification procedure is to compare lineups[1]. A recent eyewitness identification study compared the most commonly used procedure in the US to the procedure used in the UK (Seale-Carlisle & Mickes, 2016). In the

[1] Live lineups and showups (1-member identification procedures) are outside the scope of this paper.

majority of jurisdictions in the US, eyewitnesses are presented with photos of the police suspect (who is innocent or guilty) and five fillers (Police Executive Research Forum, 2014). All six photos are shown at the same time and a decision can be made at any time. In the UK (England and Wales), where the procedural guidelines are dictated by PACE Code D (1984; 2017), eyewitnesses are presented with videos of the police suspect and typically eight individuals, or fillers. Each 15-second video shows the lineup member from the shoulders up, facing forward, turning from side to side (to show left and right profile views), and then turning back to show the front view again. All nine videos are shown, one after the other, two times prior to expressing a decision. Witnesses can elect to see any lineup member(s) as many times as they would like and are offered to be shown photos of all of the lineup members together.

In any type of lineup, witnesses can make one of three possible decisions: identify the suspect, identify a filler, or reject the lineup (i.e., make no identification). Based on these decisions, there are five possible outcomes: correct identifications (ID), false IDs, correct rejections, misses, or filler IDs. A correct ID occurs when the guilty suspect in a target-present lineup is identified; a false ID occurs when the innocent suspect in a target-absent lineup is misidentified. A miss occurs when a target-present lineup is rejected; a correct rejection occurs when a target-absent lineup is rejected. A filler ID occurs when a filler is identified from either a target-present or a target-absent lineup.

In the study by Seale-Carlisle and Mickes (2016), participants first watched a video of a mock crime and were then tested using either a US or a UK lineup. They found that discriminability was significantly better when memory was tested using the US procedure. In other words, participants in the US condition made more correct IDs and fewer false IDs than those in the UK condition. Discriminability is one type of accuracy that is of interest to the

criminal justice system (Mickes, 2015). The second type of accuracy that is important to the criminal justice system is positive predictive value (PPV), an indicator of reliability. PPV measures the likelihood that the identified suspect is actually the perpetrator. The US lineup procedure also gave rise to higher reliability than the UK lineup procedure. In other words, participants who identified suspects in the US condition had greater accuracy (i.e., the identified suspects were more likely to be guilty) compared to participants who identified suspects in the UK condition. Thus, the US lineup was better for both types of accuracy.

The study by Seale-Carlisle and Mickes (2016) was designed to directly compare the US and UK procedures as they are used in practice, and we use this as a basis to decide which comparisons to make to determine which aspects of lineups will lead to better memory performance. Because there are so many differences between them, it is not clear what explains the superior performance of the US procedure. One possible candidate, as proposed by Seale-Carlisle and Mickes, is presentation format (simultaneous vs. sequential lineups). There are two competing theories that make predictions about the effect of presentation format on performance: the absolute-relative judgment theory (Wells, 1984) and the diagnostic feature detection (DFD) theory (Wixted & Mickes, 2014).

A relative decision is made when the lineup member who most resembles the memory of the perpetrator is identified. An absolute decision is made when the lineup member who matches the memory of the perpetrator is identified. Simultaneous lineups are more likely to engender relative decisions and sequential lineups are more likely to engender absolute decisions (Lindsay & Wells, 1985). According to Wells (1984, 1993) absolute decision strategies lead to lower false ID rates but similar correct ID rates (Wells, 1993). This effectively means that simultaneous lineups would yield lower discriminability than sequential lineups (Clark, Erickson, & Breneman, 2011). The DFD theory makes the opposite

prediction. It holds that the ability to make relative comparisons from a simultaneous lineup improves discriminability because non-diagnostic features can easily be appreciated and therefore discounted (Wixted & Mickes, 2014).

The main goal of the current research is to systematically isolate and compare various aspects of lineups to determine their effects on discriminability and reliability. The results will help to achieve a greater theoretical understanding of performance and may also suggest ways to improve current procedures. In five experiments, we compared (1) presentation format: simultaneous vs. sequential (Experiments 1 and 5); (2) stimulus format: photos vs. videos (Experiment 2); (3) number of views: 1-lap vs. 2-lap vs. choice in video lineups (Experiment 3a) and photo lineups (Experiment 3b); and (4) lineup size: 6- vs. 9-lineup members (Experiment 4). We next describe the general method and general analyses used for all of the experiments, followed by a presentation of the theoretical background and rationale, specific methods, and results of each experiment.

General Method and Analyses

There is substantial overlap in the methods and analyses across the five experiments. Thus, before detailing the experiments and presenting their results, the general methods and analyses are described. The data from all of the experiments are available at https://osf.io/wvm26/.

Participants

Participants were recruited from Amazon Mechanical Turk (MTurk; www.mturk.com) and were told that the researchers were interested in learning about the eyewitness experience. They took part in exchange for a small monetary sum that is standard for MTurk workers. The MTurk set up and experiment programs prevented participants from taking part

more than once. Participants were randomly assigned to a condition and a target-present or target-absent lineup. Ethical approval was granted by a Royal Holloway, University of London Ethics Committee; project number 2015/026.

Materials

With the exception of Experiment 5, the stimuli were the same as those used in Seale-Carlisle and Mickes (2016).

Video. A young adult White male acted in a 20 second video of a mock theft of a purse in an unoccupied office. The target's face was clearly shown for 8 seconds. The target's face was shown from the front for approximately 5 seconds and left and right profile views were shown for approximately 1.5 seconds each.

Lineups. A London Metropolitan Chief Inspector, trained in identification procedures, filmed the actor and selected a total of nine fillers based on PACE guidelines (1984; 2011 Code D). These guidelines stipulate that the fillers "resemble the suspect in age, general appearance, and position in life." The fillers' videos were selected from the PROMAT database (the database that the London Metropolitan Police Force uses to construct and present lineups). Each 15-second video showed a lineup member from their shoulders up first facing forward, turning from side to side, and then back to face forward. Target-present lineups contained the culprit and fillers, and target-absent lineups contained fillers. For the experiments that involved photo lineups screen shots of each lineup member's face taken from the videos (front view) were used. The positions of the target and fillers were randomized for each participant. The sizes of all of the videos and photos were the same across conditions.

Procedure

All participation occurred online (but not on mobile devices). Participants consented, took part in the study phase, distractor task, test phase, provided demographic information, and were debriefed. All of the instructions were presented on the screen in text and voiceover audio. During the study phase, participants were instructed to pay close attention to the video because they would be asked questions about it later. They then watched the video and played a 5-minute game of Tetris (the distractor task). During the test phase, their memory for the perpetrator in the video was tested on a lineup. Prior to presentation of the lineup, participants were informed that, "The person from the video may or may not be in the lineup you are about to see." The lineup members were numbered based on their position in the lineup and the number was displayed above each video or photo. The response options were not available until all of the lineup members were presented. Participants selected the number corresponding to a lineup member or selected the "not present" option from a dropdown menu. Once they made this decision, a confidence scale appeared, and they rated their confidence on an 11-point scale (0 = just guessing and 100 = absolutely certain) from a dropdown menu. Participants then answered several questions about the video, including the multiple-choice validation question, "What crime was committed?" Lastly, they provided demographic information (age, gender, ethnicity, and highest level of education), and were debriefed.

Measuring Discriminability

To measure the ability to discriminate innocent from guilty suspects, both correct ID rates and false ID rates are jointly considered. Correct ID rates are the number of guilty suspects identified divided by the total number of target-present lineups. False ID rates are the number of innocent suspects identified divided by the total number of target-absent lineups. When no innocent suspect is designated, as is the case in these experiments, false ID

rates are estimated by dividing the number of target-absent filler IDs by the number of lineup members, and dividing that value by the total number of target-absent lineups. We report overall correct and false ID rates, but do not report inferential statistics on these separately because by doing so discriminability and response bias can be confounded. Our main goal is to assess the impact of aspects of lineups on discriminability and reliability, we therefore conduct ROC analysis and CAC analysis.

Receiver operating characteristic (ROC) analysis. To conduct ROC analysis, which provides a measure of discriminability, we plotted the correct ID rates against the estimated false ID rates (obtained by dividing by lineup size) on the bottom x-axis and the target-absent filler ID rates on the top x-axis for every level of confidence. Presenting the figures this way makes it clear that that ROC data can be considered using either incorrect ID rate (e.g., Mickes et al., 2017). The points range from the highest level of confidence (the leftmost point that represents the most conservative responses) to the lowest level of confidence (the rightmost point that represents all responses, including the most liberal responses) to form the ROC curve. The higher the ROC curve, the better the ability to discriminate.

Discriminability was quantified by computing partial area under the curve (pAUC) for each condition and then statistically comparing conditions using the pROC package in R (Robin et al., 2011). When there is no designated innocent suspect (as is the case in these experiments), pAUC can be measured either by using the filler IDs from the target-absent lineups or the estimated false IDs (Wixted & Mickes, 2015a). This is legitimate because in the mind of the participant, for a fair lineup (i.e., the suspect does not stand out among the fillers) the innocent suspect is just another filler (i.e., the innocent suspects and fillers are drawn from the same distribution) (e.g., Mickes & Gronlund, 2017; Semmler, Dunn, Mickes & Wixted, 2018). By conducting pAUC analysis with the target-absent filler IDs, there is more

power to detect a difference (e.g., Mickes et al., 2017; Wilson, Seale-Carlisle & Mickes, 2018). Therefore, we use this approach in all of the experiments in which the lineup sizes were the same across conditions (all except in Experiment 4 where the estimated false IDs were used) (Mickes, 2015; Seale-Carlisle & Mickes, 2016).

To compute pAUC, a cutoff point must be set. The cutoffs used were the filler ID rate or estimated false ID rate (for Experiment 4) of the rightmost point from the condition that yielded the most conservative overall responding of the conditions being compared. For the experiments with three conditions (Experiments 3a and 3b), Bonferroni corrections were made. A tutorial of ROC analysis of lineup data is provided in Gronlund et al. (2014).

Signal detection model fits. pAUC provides an atheoretical estimate of discriminability. Because some concerns have been expressed that pAUC may not correspond to underlying (theoretical) discriminability (Lampinen, 2016; Levi, 2016; Wells, Smalarz, & Smith, 2015; Smith, Wells, Lindsay, & Penrod, 2017; Wells, Smith, & Smalarz, 2015; but see Albright, 2017; National Research Council, 2014; Rotello & Chen, 2016; Wixted & Mickes, 2015a; 2015b; Wixted, Mickes, Wetmore, Gronlund & Neuschatz, 2015), we also fit two signal detection-based models to the data. We chose the two models that map onto the two theories that predict either a simultaneous or sequential advantage: the Ensemble model (Wixted, Vul, Mickes, & Wilson, 2018) and the Independent Observations model (Macmillan & Creelman, 2005; similar to the BEST model; Clark 2003; Clark, et al. 2011), respectively.

In the Independent Observations model (Macmillan & Creelman, 2005), one Gaussian distribution represents the memory signals of innocent suspects (and the fillers for a fair lineup) and one Gaussian distribution represents the memory signals of guilty suspects. A 6-person target-absent lineup consists of six random draws from the innocent suspect distribution, and a 6-person target-present lineup consists of five random draws from the

innocent suspect distribution and one random draw from the guilty suspect distribution. An identification is made if the face in the lineup that generates the maximum memory signal exceeds the lowest of the decision criteria, otherwise no identification is made. This model assumes an absolute decision strategy (i.e., only the face that generates the strongest memory signal is considered regardless of the strength of the other faces).

The Ensemble model (Wixted et al., 2018) is the mathematical instantiation of the DFD theory (Wixted & Mickes, 2014). It is the same as the Independent Observation model except for the decision rule. In the Ensemble model, an identification is made when the difference between the strongest memory signal and the average of the memory signals generated by all of the faces in the lineup is strong enough to exceed the lowest decision criterion, otherwise no identification is made. This model assumes a relative decision strategy (i.e., the face that generates the strongest memory signal is considered relative to the strength of the other faces).

To fit the full models to the data, suspect IDs and filler IDs (from target-present and target-absent lineups) were binned into low (ratings of 0-60), medium (ratings of 70-80) and high (ratings 90-100) levels of confidence[2]. Thus, there were three decision criteria: low, medium, and high. The fixed parameters were $\mu_{lure} = 0$, $\sigma_{lure} = 1$, and $\sigma_{target} = 1$. The free parameters were μ_{target} (i.e. d', discriminability) and the three decision criteria: $_{C1}$ (low), $_{C2}$ (medium), and $_{C3}$ (high). These parameters were adjusted until the differences between observed and predicted values were minimized. We then compared the fits from the full models with the fits of the constrained models (i.e., d' was constrained to be equal across

[2] Because there are often few responses in the low confidence levels we collapse 0-60 for a low confidence bin (e.g., Mickes, 2015; Seale-Carlisle & Mickes, 2016; Wilson et al., 2018).

conditions). The curves in the ROC plots (Figures 1A, 2A, 3A, 4A, 5A, and 6A) were produced from the full Ensemble model fits.

Measuring Reliability

To measure positive predictive value (PPV) – the likelihood that the suspect identified was guilty – CAC analysis was conducted (Mickes, 2015). PPV is given by

$$PPV = \frac{CID_{conf}}{CID_{conf} + FID_{conf}/k}$$

where CID_{conf} is the number of suspect IDs made with a particular level of confidence from target-present lineups, FID_{conf} is the number of filler IDs from target-absent lineups, and k is lineup size (such that FID_{conf} / k is the estimated number of false IDs to innocent suspects). PPV is computed for each binned level of confidence (low, medium, and high), and plotted in a CAC. Standard errors are estimated using a bootstrap procedure described in Seale-Carlisle and Mickes (2016).

Relative frequencies

Our visual presentations of data within a given condition make use of symbols that vary in size to show the number of observations that they are based on, contrary to the typical approach of using symbols that are all of the same size. The size of the symbol used to represent a given data point visually illustrates how many observations contributed to the calculation of that point. Symbol size in our figures is based on the relative frequencies (RF) of each point. RF is defined by

$$RF_{conf} = \frac{CID_{conf} + FID_{conf}/k}{CID + FID/k}$$

where, again, CID_{conf} is the number of correct (i.e., suspect) IDs from target-present lineups made with a given level of confidence (*conf*), FID_{conf} is the number of filler IDs from target-

absent lineups made with the same level of confidence, and *k* is lineup size. *CID* is the total number of suspect IDs from target-present lineups and *FID* is the total number of filler IDs from target-absent lineups. As an example, consider a condition in which *CID* = 126 and *FID* = 149. Imagine that, for IDs made with 100% confidence, $CID_{100\%}$ = 16, $FID_{100\%}$ = 20, and *k* = 9 (a UK lineup). To depict the size of that point, $RF_{100\%}$ = (16 + (20/9)) / (126 + (145/9)) = 0.13. To depict the size of the point for IDs from the same condition made with 80% confidence (where $CID_{80\%}$ = 31, $FID_{80\%}$ = 7, and *k* = 9), $RF_{80\%}$ = (31 + (7/9)) / (126 + (145/9)) = 0.22. If all confidence levels were equally represented, then the points would be the same size. In this example though, of all the suspect IDs, 13% were made with 100% confidence, and 22% were made with 80% confidence. Therefore, the point for 80% confidence would be 1.7 (i.e., 0.22 / 0.13 = 1.7) times larger than the point for 100% confidence[3].

Experiment 1

In much prior research, to measure performance, the diagnosticity ratio (DR; correct ID rate/false ID rate) was used. Empirically, the DR was often found to be higher for sequential lineups (e.g., Lindsay, 1999; Steblay, et al., 2001; Steblay et al, 2011). This is why the sequential lineup was once considered to be diagnostically superior to the simultaneous lineup (e.g., Lindsay, 1999; Steblay, Dysart, Fulero & Lindsay, 2001; Steblay, Dysart & Wells, 2011; Lindsay & Wells, 1985). However, the DR is an unsuitable measure of discriminability because it conflates response bias with discriminability (Clark, Erickson & Breneman, 2011; Gronlund, Wixted, & Mickes, 2014; National Research Council, 2014; Wixted & Mickes, 2012). The sequential procedure generally yields fewer false IDs, but it also yields fewer correct IDs

[3] This example is from the simultaneous lineup in Experiment 1.

(i.e., results that are indicative of a trade-off). This pattern is consistent with conservative responding, not better discriminability, per se (Clark, 2012; Palmer & Brewer, 2012).

Although some disagree (e.g., Lampinen, 2016; Levi, 2016; Smith et al., 2017; Wells et al., 2015; Wells et al., 2015), we and others advocate using receiver operating characteristic (ROC) analysis to measure discriminability of lineup data (National Research Council, 2014; Rotello & Chen, 2016; Wixted & Mickes, 2012; Wixted & Mickes, 2018; Wixted & Mickes, 2015a, Wixted & Mickes, 2015b, Wixted et al., 2017). When ROC analysis is conducted, the sequential procedure only once yielded higher discriminability than the simultaneous procedure (Meisters, Diedenhofen & Musch, 2018), but the opposite outcome favouring the simultaneous procedure has often been observed (Andersen, Carlson, Carlson, & Gronlund, 2014, Carlson & Carlson, 2014, Dobolyi & Dodson, 2013; Gronlund, et al., 2012; Mickes, Flowe, & Wixted, 2012; Terrell, Baggett, Dasse, & Malavanti, 2017; Willing, Diedenhofen, & Musch, under review).[4] Furthermore, data from two police department field studies (in which ROC analysis was not conducted but discriminability was estimated in other ways) also show a simultaneous advantage (Amendola & Wixted, 2014; Wixted, Mickes, Dunn, Clark, & Wells, W., 2016).

To explain the advantage that simultaneous lineups often have over sequential lineups, Wixted and Mickes (2014) proposed the diagnostic feature detection (DFD) theory. According to this theory, eyewitnesses give more weight to diagnostic features (i.e., features that are unique to the suspect) and less weight to non-diagnostic features (i.e., features that are shared) when the lineup members are presented together. Recent model fitting evidence favours this explanation. When fitting the Ensemble model, the mathematical instantiation of

[4] These studies also found a DR advantage for the simultaneous procedure (failing to replicate prior work).

the DFD (Wixted & Mickes, 2014; Wixted, Vul, Mickes & Wilson, 2018), and the Independent Observations model (which uses an absolute decision rule) to the US and UK data in Seale-Carlisle and Mickes (2016), the former model provided a better fit. This suggests that the difference in discriminability between the US and UK lineup was, at least in part, based on a differential ability to compare lineup members (i.e., to make relative judgments).

Based on the DFD theory, we predict that one modification to the UK lineup – presenting the lineup members together while holding all other aspects of the procedure constant – will increase discriminability. We tested this prediction in Experiment 1. We also tested the prediction that confidence will be diagnostic of accuracy regardless of whether the lineup members are presented simultaneously or sequentially.

Method

Participants. Participants (N = 2000; 50% female, 49% male, and 1% did not state; age in years: M = 34.91; sd = 11.71; ethnicity: White 65%, Asian 21%, Black 6%, Hispanic 5%, Native American 1%, Other 2%, and did not state 1%) were randomly assigned to either the sequential (target-present n = 526 and target-absent n = 492) or simultaneous (target-present n = 483 and target-absent n = 499) lineup condition. The data from the participants who incorrectly answered the validation question (n = 88) and had internet connectivity problems (n = 5) were excluded from the analyses. Thus, the analyses were conducted on the remaining participants (n = 1907).

Materials and Procedure. The lineups were 9-member lineups (in accordance with PACE guidelines; PACE, 2017). The videos were the same size for both conditions. Target-present lineups contained the guilty suspect and 8 fillers, and target-absent lineups contained 9 fillers. In the sequential lineup condition, during the test phase, each video was

presented one-at-a-time. In the simultaneous lineup condition, during the test phase, all members' still photos were presented together in a 3 x 3 grid, and each video played one-at-a time. For both conditions, the lineup lapped twice (i.e., all lineup members' videos played two times) before a decision could be expressed.

The instructions for those in the sequential lineup condition were,

> You will see one lineup member at a time. One member will move to show you their profile. Then the next member will move, and the next, until all of them have cycled through. They will cycle through twice. If you see the person from the video, pick him. If you do not see the person from the video, pick the "not present" option.

The instructions for those in the simultaneous condition were the same, except instead of the instructions reading "You will see one lineup member at a time," they read "You will see all of the lineup members at the same time."

Results and Discussion

Table 1 shows the number of rejections, correct IDs, and filler IDs at each level of confidence for both conditions. The overall average correct ID rate in the simultaneous condition was 0.27 and 0.20 in the sequential condition (see Table A1 for the overall suspect ID rates, filler ID rates, and no ID rates for target-present and target-absent lineups per condition). The overall average estimated false ID rate in the simultaneous condition was 0.07 and 0.08 in the sequential condition. This pattern means that overall discriminability was directionally higher in the simultaneous condition. For a complete picture of discriminability differences across the family of correct and false ID rates per condition, we conducted ROC analysis (Gronlund et al., 2014; Wixted & Mickes, 2012).

As shown in Figure 1A, the sequential condition yielded a lower ROC than the simultaneous condition. With regard to atheoretical, or empirical, estimates of

discriminability, using a false ID rate cut-off of 0.63[5], the pAUC for the simultaneous condition (0.11) was significantly higher than the pAUC for the sequential condition (0.07), $D = 3.17$, $p = .002$. This result is consistent with the often-replicated differences between the US sequential photo lineups and the US simultaneous photo lineups, which generally favor the latter (Gronlund, et al., 2012; Mickes et al., 2012; Carlson & Carlson, 2014, Dobolyi & Dodson, 2013). This discriminability result also replicates the previous US vs. UK comparison (Seale-Carlisle & Mickes, 2016).

With regard to theoretical estimates of discriminability, both the Independent Observations model and the Ensemble model estimated d' to be higher in the simultaneous condition than in the sequential condition (see Table A2). The Independent Observations model provided a tolerable fit, $\chi^2(10) = 16.09$, $p = 0.097$, and the Ensemble model provided a slightly better fit, $\chi^2(10) = 15.54$, $p = .114$. The difference in d' between conditions was significant because constraining d' in the Ensemble model to be equal between conditions resulted in a poorer fit, $\chi^2(11) = 23.01$, $p = 0.138$, and the difference between the full and constrained Ensemble fits was highly significant, $\chi^2(1) = 7.47$, $p = 0.006$. Thus, as in previous studies (Colloff, Wade, & Strange, 2016; Colloff, Wade, Wixted, & Maylor, 2017; Seale-Carlisle & Mickes, 2016; Wilson et al., 2018), measures of empirical and theoretical discriminability are consistent. The consistency between atheoretical and theoretical measures of discriminability is important to check because it is not a guaranteed outcome (Lampinen, 2016; Smith et al., 2017; Wixted & Mickes, 2018).

Figure 1A also shows the relative frequencies of each level of confidence. In both conditions, very few identifications were made with the lowest levels of confidence (i.e., 0-

[5] The scale on the x-axis is the estimated false ID rates, and pAUCs were measured using all of the filler IDs from the target-absent lineups.

40% confident). This is standard in our experience, and why we often collapse the confidence levels down to 0-60 for a low confidence bin (e.g., Mickes, 2015). In the simultaneous condition, more IDs were made with medium confidence (70-80%) than were made with high or low confidence. In the sequential condition, responding was more distributed throughout the medium and high levels of confidence and more IDs were made with low levels of confidence.

As shown in Figure 1B, for both conditions, identifications made with high confidence had higher PPV than identifications made with medium confidence, which, in turn, had higher PPV than identifications made with low confidence. Aside from a difference between conditions in the medium confidence levels, the conditions did not differ appreciably in PPV. For both conditions, as shown by the relative size of the points, there were fewer high confidence responses than medium or low confidence. For the sequential condition, the majority of identifications were made with low confidence, and for the simultaneous condition, the majority of identifications were made with medium confidence.

Figure 1B also shows that accuracy was somewhat lower for the highest levels of confidence in the sequential condition than the simultaneous condition, but that difference was not significant. Identifications made with high confidence in either condition were more likely to be accurate than at lower levels of confidence. This pattern is consistent with previous research (e.g., Wixted & Wells, 2017), though high-confidence accuracy was noticeably lower with this stimulus set than is usually observed (e.g., > 95% accuracy). We found the same result using this stimulus set previously (Seale-Carlisle & Mickes, 2016), and it may indicate that some of the fillers were too similar to the target to achieve higher levels of accuracy.

Experiment 2

It makes intuitive sense that videos over photos would facilitate memory performance because videos simply contain more information than photos. Does other extant empirical evidence support the intuition that performance is better when video versus photo stimuli are used? In list-learning recognition memory experiments in which participants study a list of items (e.g., scenes, faces) and then are tested on those items and lure items, discriminability is better for moving images than for static images (e.g., Goldstein, Chance, Hoisington, & Buescher, 1982). The size of the effect is typically largest when the items at study and test are matched (i.e., moving at study and moving at test; Buratto, Matthews, & Lamberts, 2009). Whether this applies to richer, more real-life stimuli is unclear.

In forensically-relevant recognition memory experiments (in which participants watch a mock crime and then are tested on a lineup), comparisons of moving vs. static stimuli have yielded mixed results. Valentine, Darling, and Memon (2007) compared sequential video lineups to sequential photo lineups. We computed *d'* from the correct ID rate and false ID rate that they reported (using the formula recommended by Mickes, Moreland, Clark & Wixted, 2014), and found that discriminability was better for video lineups than photo lineups (collapsed across their existing and strict rules conditions, average d'_{video} = 1.56 and average d'_{photo} = 0.55).

Another experiment, conducted by the same researchers (Darling, Valentine & Memon, 2008), found the opposite. Here, discriminability was better for sequential photo lineups (collapsed across match to suspect and match to description conditions, average d'_{photo} = 1.44 and average d'_{video} = 1.20). To get an indication of the impact that the photos and videos had on reliability, we computed DR. The DR, while it is not a measure of discriminability, is a measure of PPV when the base rates of target-present and target-absent

lineups are equal (Mickes, 2016). The same inconsistent pattern emerged in that DR was higher for photo lineups in Darling et al., but lower in Valentine et al. (2007).

In those prior studies, confidence was either not collected or not reported in a way that ROC analysis and CAC analysis[6] could be conducted. Additionally, in those prior studies, the lineups were all sequentially presented. Never before (to our knowledge) has this stimulus format been compared in simultaneous lineups. Therefore, in Experiment 2, we compare the simultaneous video lineup with a simultaneous photo lineup. This experiment allows us to test whether video presentation has any leverage over photo presentation in simultaneous lineups.

Method

Participants. Participants (N = 2275; 53% female and 47% male; age in years: M = 34.19; sd = 11.70; ethnicity: White 63%, Asian 20%, Black 7%, Hispanic 5%, Native American 1%, Other 2%, and did not state 2%) were randomly assigned to either the photo (target-present n = 604 and target-absent n = 553) or video (target-present n = 548 and target-absent n = 569) condition. The data from participants who incorrectly answered the validation question (n = 83) and had internet connectivity problems (n = 3) were excluded from analyses. Thus, analyses were conducted on the data from remaining participants (n = 2189).

Materials and Procedure. In the video condition, participants were tested on the same simultaneous video lineup as in Experiment 1. In the photo condition, participants were tested on a simultaneous photo lineup. The photos were still images taken from the videos. In both conditions, lineup members were shown in a 3 x 3 grid and a decision could be made

[6] CAC analysis is a more complete measure of PPV than computing the DR on the overall correct and false ID rates.

immediately and a response made after the videos played and after the photos were presented.

Results and Discussion

Table 2 shows the number of rejections, false IDs, correct IDs, and filler IDs at each level of confidence. The overall average correct ID rate in the video condition was 0.27 and 0.24 in the photo condition (see Table A1 for the suspect ID rates, filler ID rates, and no ID rates for target-present and target-absent lineups per condition). The overall average estimated false ID rate in the video condition was 0.07 and 0.05 in the photo condition. Whether there is a shift in responding in the conservative direction when memory is tested using photos can be informed by ROC analysis.

Figure 2A shows the ROC curves for both photo and video conditions. Using a false ID rate cutoff of 0.52, the pAUC was marginally higher in the photo condition (0.072) than the video condition (0.055), $D = 1.91$, $p = 0.056$. Consistent with the ROC analysis, the fits of the Independent Observations and Ensemble models also estimated d' to be higher in the photo condition than in the video condition (see Table A2). The full Independent Observations model provided a poor fit, $\chi^2(10) = 21.76$, $p = 0.016$. The full Ensemble model provided a noticeably better fit, $\chi^2(10) = 15.82$, $p = 0.105$. Constraining d' in the Ensemble model to be equal across conditions resulted in a similar fit, $\chi^2(11) = 16.53$, $p = 0.123$, and the fits between the full and constrained Ensemble models were not significantly different, $\chi^2(1) = 0.71$, $p = 0.399$. Thus, overall, there was a trend towards higher discriminability for the static photo condition, but, if the effect is real, it is not very strong. On the other hand, the data are incompatible with the notion that moving images increase discriminability. If anything, it is the other way around.

Overall responding was more conservative in the photo condition than in the video condition, which is clear in Figure 2A. That is, the points in the former condition are shifted leftward relative to the latter condition. Figure 2A also shows relative frequencies of each level of confidence. Again, in both conditions, very few identifications were made with the lowest levels of confidence (0-30% confidence). In both conditions, responses were mostly, and fairly evenly, distributed throughout the medium confidence levels.

Figure 2B shows the CAC plot. The larger points indicate that more participants, in both conditions, made IDs with low confidence (0-60% confidence) than with medium and high confidence. PPV was higher for the photo condition for the IDs made with medium and high levels of confidence.

The comparison of video lineups with photo lineups revealed no significant difference in discriminability. Although it does not appear that the dynamic nature of the video lineup aided discriminability (in fact, the photo lineup ROC fell on a slightly higher curve), it is possible that the brief presentation of the target's profile may have accounted for a lack of a benefit for the video lineup (which showed front and profile views). Indeed, the dynamic superiority effect is most noticeable when the study and test items show the same view (Buratto et al., 2009). Once again when identifications are made with high confidence, they are more likely to be accurate than identifications made with low confidence, showing again that confidence has informative value about accuracy.

Experiment 3a

The standard procedure in the UK requires an eyewitness to lap through the lineup twice before the decision is expressed. Does lapping through the lineup twice affect discriminability or reliability? Some of the previous studies on the topic included sample sizes

that were too small and/or obtained discriminability scores that were too low (Lindsay, Lea, & Fulford, 1991; MacLin & Phelan, 2007; Steblay, et al., 2011 Experiment 1). In Experiment 2 of Steblay et al., participants in a forensically-relevant experiment were assigned to one of two conditions (required to lap twice or chose to lap – once or twice). We computed *d'* from the reported correct ID rates and estimated false ID rates of the final decision, and *d'* (and DR) was higher for those in the elected lap condition (collapsed across lapping once and twice).

In another forensically-relevant experiment, participants were required to lap twice or were given the choice to do so after viewing the first lap (Horry, Palmer & Brewer, 2015). Similar to Steblay et al. (2011), for each lineup member, in each lap, participants responded yes or no. We computed *d'* from their correct ID rates and estimated false ID rates of the final decision, and *d'* was higher (but not significantly) for those in the choice condition who opted to view the lineup only once compared to those in the same condition who opted to lap twice and those in the required lap twice condition. The DR was also higher in the former than the latter groups.

In those previous experiments, a stopping rule was employed. The standard UK lineup procedure requires eyewitnesses to lap through the lineup twice (at least) before a decision is expressed. Because our focus is on the effect that lapping may have on discriminability and reliability on the UK lineup, participants in this experiment (and in Experiment 3b) expressed their decision after they lapped through the lineup, either once or twice depending on the condition, and rated their confidence.

Method

Participants. Participants (*N* = 3105; 58% female and 42% male; age in years: *M* = 30.79; *sd* = 12.02; ethnicity: White 53%, Asian 30%, Black 5%, Hispanic 7%, Native American

1%, Other 2%, and did not state 1%) were randomly assigned to the 1-lap condition (target-present *n* = 508 and target-absent *n* = 535), 2-lap condition (target-present *n* = 502 and target-absent *n* = 523), or lap choice condition (target-present *n* = 534 and target-absent *n* = 503). The data from participants who incorrectly answered the validation question (*n* = 72) and those who had internet connectivity problems (*n* = 3) were excluded from analyses. Thus, the analyses were conducted on the remaining participants (*n* = 3030).

Materials and Procedure. All lineups were video lineups presented sequentially. Participants expressed their decision after seeing all of the lineup members one time (1-lap condition or choice condition) or twice (2-lap condition or choice condition). For participants in the choice condition, they were given the option to lap one more time after they saw the first lineup lap.

Results and Discussion

Table 3 shows the number of rejections, false IDs, correct IDs, and filler IDs at each level of confidence. The overall average correct ID rates for the 1-lap, 2-lap, and choice conditions were 0.25, 0.24, 0.25, respectively (see Table A1 for the overall suspect ID rates, filler ID rates, and no ID rates for target-present and target-absent lineups per condition). The overall average estimated false ID rates for the 1-lap, 2-lap, and choice conditions were 0.08, 0.08, 0.07, respectively.

For those in the choice condition who opted for only 1 lap (90%), the overall average correct ID rate was 0.25 and 0.20 for those who opted for 2 laps. The overall average estimated false ID rate for those in the choice condition who opted for only 1 lap was 0.08 and 0.07 those who opted for 2 laps. Because there were too few observations to conduct ROC analysis (because so few of those in the choice condition opted for a second lap), we compared *d'* values using the G statistic (Mickes, Moreland, Clark & Wixted, 2014).

Discriminability was somewhat better for participants who opted for one lap (d' = 0.73) compared to participants who opted to view the lineup twice (d' = 0.64), but this difference was not significant (p = 0.731).

Figure 3A shows the ROC curves for all of the lapping conditions. Using a false ID rate cutoff of 0.31[7], pairwise comparisons, using Bonferroni corrections, revealed no significant differences in pAUC values: 1-lap (0.11) vs. 2-lap (0.09), D = 1.13, p = 0.267; 1-lap vs. choice (0.10), D = 0.24, p = 0.810; choice vs. 2-lap, D = 0.87, p = 0.374. The Independent Observations and Ensemble models estimated similar d' values across the three conditions: d' was lowest for the 2-lap condition and highest for the 1-lap condition (see Table A2). Thus, whether comparing theoretical discriminability estimated by one of these two models or empirical discriminability measured by conducting ROC analysis, the conclusion is the same. The full Independent Observations model fit was tolerable, $\chi^2(15) = 22.28$, p = 0.101, and the full Ensemble fit was much better, $\chi^2(15) = 9.66$, p = 0.841. Constraining d' in the Ensemble model to be equal across conditions also resulted in a good fit, $\chi^2(17) = 10.32$, p = 0.799, which was not significantly different than the full Ensemble model fit, $\chi^2(2) = 0.66$, p = 0.719.

Figure 3A shows the relative frequencies of each level of confidence. In both conditions, like in Experiments 1 and 2, very few identifications were made with the lowest levels of confidence. In the 1-lap condition, responses were mostly, and fairly evenly, distributed throughout the medium confidence levels, and in the other two conditions, many responses were made with 70% confidence.

As shown in Figure 3B, all conditions yielded similar PPV values (with the 2-lap condition yielding a slightly, but not significantly, lower PPV for the IDs made with the highest

[7] The scale on the x-axis is the estimated false ID rates, and pAUCs were measured using all of the filler IDs from the target-absent lineups.

and medium levels of confidence). In all conditions, the fewest responses were made with high confidence, which is another indication that the stimuli used in this experiment made the task hard for the participants. Typically, for example, there are far more high-confidence decisions than low-confidence decisions.

Experiment 3b

We investigated whether lapping once, twice, or making a choice to lap once or twice in photo lineups would yield similar results to Experiment 3a. In that experiment, lapping through lineup members' videos once, twice, or having a choice to lap or not had no differential effect on discriminability and reliability. We predicted that the pattern would be no different with photo lineups.

Method

Participants. Participants (N = 3003; 45% female and 55% male; age in years: M = 34.70; sd = 11.35; ethnicity: White 53%, Asian 32%, Black 5%, Hispanic 5%, Native American 2%, Other 2%, and did not state 1%) were randomly assigned to either the 1-lap condition (target-present n = 547 and target-absent n = 487), the 2-lap condition (target-present n = 489 and target-absent n = 498), or lap choice condition (target-present n = 509 and target-absent n = 473). The data from participants who answered the validation question incorrectly were excluded from the analyses (n = 110).

Materials and Procedure. Everything was the same as in Experiment 3a except photos were used in place of videos.

Results and Discussion

Table 4 shows the number of rejections, false IDs, correct IDs, and Filler IDs at each level of confidence. The overall average correct ID rates for 1-lap, 2-lap, and choice conditions were 0.27, 0.28, and 0.29, respectively (see Table A1 for the overall suspect ID rates, filler ID rates, and no ID rates for target-present and target-absent lineups per condition). The overall average estimated false ID rates was 0.08 for the 1-lap, 2-lap, and choice conditions. Figure 4A shows the relative frequencies of each level of confidence and the pattern is similar to Figure 3A.

For those in the choice condition who opted for only 1 lap (82%), the overall average correct ID rate was 0.25 and 0.20 for those who opted for 2 laps and the overall average estimated false ID rate for those in the choice condition who opted for only 1 lap was 0.08 and 0.07 for those who opted for 2 laps. Discriminability was better for participants who opted for one lap ($d' = 0.85$) compared to participants who opted to view the lineup twice ($d' = 0.76$), but this difference, as in Experiment 3a, was not significant ($p = 0.741$).

Figure 4A shows the ROC curves for all of the lapping conditions. Using a false ID rate cutoff of 0.30[8], pairwise comparisons, using Bonferroni corrections, revealed no significant differences in pAUC values: 1-lap (0.11) vs. 2-lap (0.10), $D = 0.92$, $p = 0.365$; 1-lap vs. choice (0.11), $D = 0.03$, $p = 0.974$; 2-lap vs. choice $D = 0.95$, $p = 0.355$. The Independent Observations and Ensemble models (see Table A2) estimated similar d' values across the three conditions. For both models, d' was lowest for the 2-lap condition and highest for the choice condition. The full Independent Observations model provided a poor fit, $\chi^2(15) = 37.24$, $p = 0.001$. The full Ensemble model fit was better, but the fit was still poor, $\chi^2(15) = 30.54$, $p = 0.010$. This is likely due to a poor fit of the choice condition, $\chi^2(5) = 14.89$, $p =$

[8] The scale on the x-axis is the estimated false ID rates, and pAUCs were measured using all of the filler IDs from the target-absent lineups.

0.011, as the fit for the 1-lap, $\chi^2(5) = 1.43$, $p = 0.921$, and 2-lap conditions, $\chi^2(5) = 3.42$, $p = 0.636$, were good. The fit when d' was constrained to be equal was also poor, $\chi^2(17) = 30.83$, $p = 0.021$, and the fits between the full and constrained Ensemble model were not significantly different, $p = 0.865$.

Figure 4B shows that the PPV pattern was the same as in Experiment 3b. At every level of confidence, PPV was lower for those in the 2-lap condition. Across conditions, there were fewer IDs made with high confidence than medium or low confidence, further confirming that from the participants' point of view, this was a difficult task (perhaps accounting for why high-confidence accuracy was unusually low for this set of stimuli).

Experiment 4

Another notable difference between the standard US and UK lineups is that the former has 6-members whereas the latter has 9-members. Did the smaller lineup size account for the better performance when memory was tested on the US lineup (Seale-Carlisle & Mickes, 2016)? In an experiment of lineup size, participants were assigned to a 4-, 8-, 12-, 16-, or 20-person lineup[9] (Nosworthy & Lindsay; 1990; Experiment 2). We computed d' and DR values on the reported correct ID rates and false ID rates (computed by using the most identified filler). There was some variation in discriminability (d' = 1.18, 1.04, 1.18, 1.04, 0.70 for 4-, 8-, 12-, 16-, or 20-person lineups, respectively) and DR (4.36, 4.71, 4.36, 4.71, 2.47 for 4-, 8-, 12-, 16-, or 20-person lineups, respectively). Some of that variation is likely due to the fact that there were only 54 participants per condition. Other studies in which lineup sizes were measured also had sample sizes that were too small to determine the effect

[9]Whether these were sequential or simultaneous lineups is unclear.

that size of lineups has on discriminability and reliability (Lindsay, Smith, & Pryke, 1999; Pozzulo, Dempsey, & Wells, 2010).

In another lineup size experiment, participants studied eight targets and were presented with 16 lineups of different sizes (Meissner, Tredoux, Parker & Maclin, 2005, Experiment 2). Discriminability when memory was tested on sequential and simultaneous lineups as measured using *d'* showed less variability (*d'* = 1.69, 1.57, 1.58, 1.52, 1.50, 1.58 for 2-, 4-, 6-, 8-, 10-, 12- person lineups, respectively). Discriminability was highest for showups, which is inconsistent with other literature (e.g., Mickes, 2015; Wetmore et al., 2015). This may be due to the fact that this was a within-subjects design (i.e., all participants were tested on 16 lineups) versus the 1-trial per participant forensically relevant design. In within-subjects tasks, participants may change strategies as they proceed through the list. While the pattern of discriminability across lineup sizes was unclear, the pattern of DR was much clearer, where the DR increased as lineup size increased (5.21, 6.56, 8.50, 9.0, 10.0, 12.67 for 2-, 4-, 6-, 8-, 10-, 12- person lineups, respectively).

Whether this pattern holds when a forensically relevant design is used is unknown. We therefore investigated this in Experiment 4 by comparing performance on lineups that contained either 6 or 9 members. The lineup procedure followed the standard procedure used in the UK with the exception of, in one condition, participants viewed 6, not 9, lineup members.

Participants. Participants (*N* = 2015; 48% female, 51% male, 1% did not state; age in years: *M* = 33.62; *sd* = 11.09; ethnicity: White 57%, Asian 26%, Black 7%, Hispanic 6%, Native American 1%, Other 2%, and did not state 1%) were randomly assigned to either the 6-member condition (target-present *n* = 515; target-absent *n* = 497) or 9-member condition (target-present *n* = 528; target-absent *n* = 475). The data from participants (*n* = 77) who

incorrectly answered the validation question were excluded from analyses. Thus, analyses were conducted on the remaining participants (n = 1938).

Materials and Procedure. In both conditions, video lineups were presented sequentially, following the standard UK lineup protocol. For the 6-member condition, in the target-present lineups 5 of the 9 fillers were randomly selected for each participant, and in the target-absent lineups 6 of the 9 fillers were randomly selected for each participant.

Results and Discussion

Table 5 shows the number of rejections, false IDs, correct IDs, and filler IDs at each level of confidence. The overall average correct ID rate and estimated false ID rate for the 6-member lineup was 0.33 and 0.11, respectively (see Table A1 for the overall suspect ID rates, filler ID rates, and no ID rates for target-present and target-absent lineups per condition). The overall average correct ID rate and estimated false ID rate for the 9-member lineup was 0.24, 0.07, respectively.

Figure 5A shows the ROC curves for both photo and video conditions. To compare different lineup sizes, the estimated false IDs were used. Using a false ID rate cutoff of 0.92, the pAUC was slightly, but not significantly, higher in the 9-member condition (0.011) than the 6-member condition (0.010), D = 0.67, p = 0.446. Consistent with the ROC analysis, the Independent Observations model estimated slightly higher d' values for the 9-member condition (see Table A2). The Ensemble model however estimated slightly higher d' for the 6-member condition instead. Neither model provided a good fit to these data. The full Ensemble fits were poor, $\chi^2(10) = 20.61$, p = .024, and fits constraining d' to be equal were also poor, $\chi^2(11) = 20.80$, p = 0.038. However, the fits of the full model and the constrained model were not significantly different, $\chi^2(1) = 0.19$, p = 0.663. Figure 5A shows relative

frequencies of each level of confidence, and as in the previous experiments, few identifications were made with the lowest levels of confidence (rightmost points on the ROC).

Figure 5B shows the CAC plots. Fewer IDs were made with high confidence than medium or low confidence in both conditions. As in the previous experiments, low confidence had lower PPV than the identifications made with medium and high confidence. The identifications made with medium and high confidence in the 6-member condition were lower than those identifications in the 9-member condition.

Experiment 5

In the preceding series of experiments, we isolated various factors of different lineups to assess differences in discriminability and reliability. In all of the experiments presented, the only difference was that the simultaneous, not sequential, presentation yielded better discriminability. We opted to use the same guilty suspect and fillers in Experiments 1-4 to keep the experiments as controlled as possible. Though this is consistent with all of the sequential versus simultaneous experiments published since 2012, it could be argued that the present results were specific to the set of stimuli we used. Moreover, it seems clear that these stimuli made the lineup tasks harder than is true of most lineup experiments (where many more decisions are made with high confidence and where high-confidence accuracy is noticeably higher than it was here). Therefore, in Experiment 5, we conducted a replication of Experiment 1 using another set of stimuli.

Method

Participants. Participants (N = 2019; 49% female and 51% male; age in years: M = 33.53; sd = 10.85; ethnicity: White 53%, Asian 32%, Black 6%, Hispanic 5%, Native American 1%, Other 2%, and did not state 1%) were randomly assigned to either the sequential

condition (target-present *n* = 490 and target-absent *n* = 497) or the simultaneous condition (target-present *n* = 512 and target-absent *n* = 520). Participants (*n* = 296) who incorrectly answered the validation question were excluded from analyses. The following analyses were conducted on the remaining participants (*n* = 1723).

Materials. The materials included a video and lineups.

Video. A young White male actor acted in a 21 second video of mock crime in which he sprayed graffiti on the side of a building. The target's face was clearly shown for 7 seconds.

Lineups. A different London Metropolitan Chief Inspector than the one who assisted in the stimuli creation for Experiments 1-4, also trained in identification procedures, assisted with the stimuli creation for this experiment. For the lineups, the actor was filmed, and nine filler videos were selected from the PROMAT database (based on PACE code regulations; PACE, 2011). These videos were same in action and duration as the videos used in Experiments 1-4.

Procedure. The procedure is exactly the same in Experiment 1.

Results and Discussion

Table 6 shows the number of rejections, false IDs, correct IDs, and filler IDs at each level of confidence. The average overall correct ID rate when memory was tested on a simultaneous lineup was 0.40 and 0.35 on a sequential lineup (see Table A1 for the overall suspect ID rates, filler ID rates, and no ID rates for target-present and target-absent lineups per condition). The average overall estimated false ID rate was 0.07 for both lineup conditions.

As shown in Figure 6A, the sequential condition yielded a lower ROC than the simultaneous condition. Using a false ID rate cut-off of 0.41, the pAUC for the simultaneous

condition (0.15) was higher, but not significantly, than the pAUC for the sequential condition (0.13), $D = 1.14$, $p = 0.261$. Once again, the Independent Observations model and the Ensemble model estimated d' to be higher in the simultaneous lineup condition than the sequential lineup condition (see Table A2), consistent with Experiment 1. The full Independent Observations fit was poor, $\chi^2(10) = 25.74$, $p = 0.004$, but the full Ensemble fit was good, $\chi^2(10) = 7.99$, $p = 0.630$. When d' in the Ensemble model was constrained to be equal across conditions the fit was worse, $\chi^2(11) = 10.21$, $p = 0.512$. However, the difference in the full and constrained Ensemble fits, $\chi^2(1) = 2.22$ was not significant, $p = 0.136$. Figure 6A shows the relative frequencies of each level of confidence, and as in the previous experiments, few identifications were made with the lowest levels of confidence. Now, however, more responses were made with high confidence, and the accuracy of high-confidence IDs exceeded 90% correct. This result suggests that the stimuli used for this experiment made the task easier than the stimuli used in the other experiments.

Figure 6B shows the CAC plots. The conditions did not differ at the medium and high level of confidence. More identifications were made with high and medium confidence than with low confidence for both conditions.

General Discussion

A series of experiments was conducted in order to gain a greater theoretical understanding of lineup performance that can in turn guide ways to improve lineups. Our focus was placed on differences in discriminability and reliability because these findings are important for 1) theoretical advancement and 2) policymakers and triers of truth. The approach that we took was to separate and measure the component parts of commonly used lineups. Most of the comparisons – videos vs. photos, lapping once vs. lapping twice vs. choosing to lap, and 6 vs. 9 lineup members – did not show appreciable differences. However, one comparison did: lineups that are presented simultaneously yielded better discriminability than lineups that are presented sequentially.

Simultaneous vs. Sequential Presentations

This finding replicates recent studies that found a simultaneous superiority effect and is precisely what the DFD theory predicts (Wixted & Mickes, 2014), but not what the absolute-relative theory predicts. According to the DFD theory, by presenting the lineup members simultaneously, comparisons across individuals (i.e., relative judgments) can be made. These comparisons make it readily apparent that some features are non-diagnostic (e.g., age, ethnicity, hair colour, etc.), and thus focus is placed on more diagnostic, non-overlapping features. By presenting the lineup members sequentially, shared features are less apparent to the witness. Under such conditions, non-diagnostic features are given more weight, which consequently reduces discriminability.

In accordance with this theory, the Ensemble model provided a better fit to the data in each experiment than the Independent Observations model (Experiment 4 was the only exception). The fact that the Ensemble model, which assumes a type of relative decision strategy, provided the best fits to the lineup data suggests that lineups ought to be designed to maximize relative instead of absolute judgments. Moreover, in all fits of the Ensemble

model, the conclusions made by comparing the full model and the constrained model matched the conclusions made by conducting ROC analysis and comparing pAUC. Whenever there was a significant difference in pAUC (i.e., Experiment 1), constraining *d'* to be equal across conditions provided a significantly worse fit than when *d'* was free to vary. Whereas, whenever there was no significant difference in pAUC (i.e., Experiments 2-5), constraining *d'* to be equal across conditions provided an equally adequate fit as when *d'* was free to vary. Thus, although it is possible that atheoretical discriminability measures such as pAUC and underlying discriminability measures (such as *d'*) can yield different outcomes (Lampinen, 2016), the outcomes are often consistent (Mickes et al., 2014; Rotello & Chen, 2016; Wixted & Mickes, 2015a).

In the earlier study by Seale-Carlisle and Mickes (2016), making direct comparisons between the US and UK lineups as they are typically implemented precluded conclusions about why the US lineup outperformed the UK lineup. Seale-Carlisle and Mickes (2016) suggested that it could be due to the simultaneous presentation of the US lineup, based on the prediction of the DFD theory. They also considered the possibility that participants in the UK lineup condition may lose attention during the course of the procedure because they cannot express a decision until after the lineup lapped through twice, whereas a decision could be immediately expressed in the US lineup condition.

Fitzgerald, Price, and Valentine (2018) favoured the idea that the US lineup outperformed the UK lineup because of the difference in duration, not because the simultaneous presentation is superior to the sequential presentation. They argued that the simultaneous superiority explanation is likely wrong and offered several different reasons why. One concern they expressed was that, "the length of the UK lineup procedure in this study (reportedly, ~6 mins) may have been too long for undergraduate students to attend to

the videos in an unsupervised environment." (p. 18). The results of Experiments 1 and 5 directly address this concern. Both the simultaneous and sequential lineups took the same duration (i.e., both procedures lapped twice before a decision could be expressed). And the simultaneous lineups still outperformed the sequential lineups. Though the difference was not significant in terms of pAUC in Experiment 5, the effect was in the same direction (i.e., the simultaneous lineup yielded better discriminability), which is a replication (Wilson & Wixted, 2018). These findings are consistent with the idea that simultaneous presentation leads to better discrimination ability.

Figure 7 provides further support for this idea. To construct the forest plot in Figure 7, we plotted pAUC differences ($pAUC_{simultaneous} - pAUC_{sequential}$) from all of the studies[10] that reported discriminability comparisons of simultaneous vs. sequential presentations (i.e., ROC analysis was conducted)[11]. If the points fall on 0, then there is no difference between simultaneous and sequential presentations. Of the 12 experiments, 11 reported measurements are positive (i.e., they fall to the right of 0), which shows that simultaneous presentation yields better discriminability then sequential presentation. The trend with sample size shows a clear improvement in statistical uncertainty while still supporting better discriminability for simultaneous presentation. The weighted average difference of pAUC between simultaneous and sequential presentations is 0.0103 and the weighted average 95% confidence intervals are 0.0064 – 0.0142. It is possible that, for whatever reason, these experiments somehow advantaged simultaneous lineups and that the effect is due to some

[10] We included all one trial per participant designs, but it is worth noting that Dobolyi and Dodson (2013) also reported a significant simultaneous advantage using a multi-trial per participant design.

[11] Terrell et al. (2017) did not provide confidence intervals and did not respond to our request for the data. The pAUC was higher in simultaneous condition vs. the sequential condition, but the difference was not significant.

shared methodological commonality across these experiments. However, there is theoretical reason to suggest that this is not the case.

We argue that this simultaneous advantage in discriminability arises because when the lineup members are presented at the same time, then participants can more easily discount non-diagnostic features and use diagnostic features to contribute to their decision (based on the DFD theory). The extent to which differing attentional demands play a role is an interesting question for future research.

Another concern expressed by Fitzgerald et al. (2018) was that the pattern of responding differed from other simultaneous versus sequential comparisons in Seale-Carlisle and Mickes (2016). They computed choosing rates from our experiments, and reported that choosing, including fillers, from both target-absent and target-present lineups was higher for sequential versus simultaneous lineups. This is the opposite pattern than one would normally expect because sequential lineups usually result in lower choosing rates. However, this line of reasoning overlooks two key issues.

First, UK lineups do not have a stopping rule as they do in the sequential lineup work that they cited, and it is the stopping rule that contributes to the "conservative" responding seen with the sequential procedure. Indeed, the results from Experiments 1 and 5 here, neither of which used the stopping rule, replicated the pattern of results in Seale-Carlisle and Mickes (2016). That is, choosing rates were higher for the sequential lineups than for the simultaneous lineups. The fact that the UK lineup and sequential lineup procedures differ in terms of response bias does not imply that they are not both impaired in terms of discriminability, and for the same reason. For example, both lineup procedures may suffer from the failure to discount non-diagnostic features regardless of their effects on response bias.

Second, in computing choosing rates, Fitzgerald et al. (2018) did not take into account the fact that the UK lineup has 9 members and the US lineup has 6 members. More plausible lineup members would lead to more choosing (e.g., Nosworthy & Lindsay, 1990). The results from Experiment 4, where comparisons were made from 6- and 9-member lineups, show that this is the case. Choosing rates were higher for the larger size lineup than the smaller size lineup. Combining different lineup sizes and no stopping rule, as was done in the US vs. UK comparison (Seale-Carlisle & Mickes, 2016), could explain the differences in choosing rates.

Photo vs. video lineups

Fitzgerald et al. (2018) reviewed the literature on video lineups vs. photo lineups and concluded that, at this point, "the empirical literature provides no compelling evidence in favor of either photo or video lineups." (p. 22). As Fitzgerald et al. noted, those studies that they reviewed in which a video advantage was found were underpowered. Our experiments had much greater power, and our results do not support the notion that video presentation is better in terms of discriminability. It is possible, however, that if the duration of the profile views during the initial presentation were longer, performance in the video condition may have been better. Further investigations are needed to address that possibility.

Lapping

In the UK lineup procedure, the lineup members are all viewed two times. The fact that the identification decision can be expressed only after lapping through the lineup twice, should help rule out non-diagnostic features of the lineup members, according to the DFD theory (Wixted & Mickes, 2014). This is because participants have an extra opportunity to see all of the faces, and therefore make it more likely that they notice the overlapping features

that should be discounted. However, there were no appreciable differences in discriminability for participants who lapped once or twice through the lineup. This does not necessarily count against the DFD theory, because it is possible that the second lap caused interference which may have negated any benefit seeing the lineup members twice would have otherwise had.

Another puzzle regarding the results of Experiments 3a and 3b was that, unlike in Horry et al. (2015), many fewer participants opted for the second lap. In that study, participants took part in the field, whereas in our study participants took part online. This may account for the fact that nearly 50% declined to lap twice in their study versus our study in which 86% declined to lap twice (averaging across Experiments 3a and 3b). Also, in their lineup procedure, participants could express decisions at any point whereas with our lineups, participants had to wait until all lineup members were viewed before they could express a decision. Despite these differences, in Horry et al., and our two experiments (Experiments 3a and 3b), *d'* was slightly, but not significantly, higher for those who declined to lap twice compared to those who opted to lap twice[12].

Reliability

Regarding reliability, the major finding throughout this series of experiments, consistent with previous literature, is that confidence is diagnostic of accuracy. That is, in general, the pattern indicated that high confidence identifications were higher in accuracy than medium confidence identifications (with a few exceptions), and medium confidence identifications were higher in accuracy than low confidence identifications. One puzzling result that is consistent throughout this set of experiments (and replicates the pattern in

[12] This comparison was of the *d'* values of those who declined the second lap and the second lap (but not the first lap) of those who opted to lap twice (Horry et al., 2015).

Seale-Carlisle & Mickes, 2016) is that high confidence accuracy is lower than predicted based on a body of literature that shows identifications that are made with high confidence are *very high* in accuracy. For example, averaged across PPV (and across conditions) in nine studies that used comparable scaling (i.e., 100-point confidence scale), high confidence identifications were 97% accurate (Wixted & Wells, 2017). This is true even when suboptimal variables prevail during encoding. These lab-based results accord with the real-world finding that for the DNA exoneration cases where information about the initial ID was available, none were made with high confidence (Garrett, 2011). Instead, as would be expected based on the lab data, they were all (error-prone) low-confidence IDs. However, averaged across PPV (and across conditions) in the current six experiments, high confidence identifications were only 84% accurate[13]. It may be the case that prior work mostly involved relatively good discriminability performance even for the poor encoding conditions (e.g., Semmler et al., 2017) but some aspect of our task pushed overall discriminability down and affected high confidence accordingly.

A possible explanation could be that encoding conditions did not vary across participants in the current experiments. When encoding conditions vary (like they do in the real world), the confidence-accuracy relationship becomes much stronger (Lindsay, Read, & Sharma, 1998). Thus, perhaps the actual surprise is how high PPV usually is for high confidence identifications even when encoding conditions are not varied (e.g., Wixted & Wells, 2017). Our findings might indicate a limitation of that result. For example, it might be that when discriminability is relatively low (as in Experiments 1-4) and when encoding conditions do not vary (as in Experiments 1-4), high confidence PPV may not be high. The

[13] In the current set of experiments, though high confidence PPV is lower than expected, there are, in general, fewer of those IDs than IDs made with medium and low confidence (as shown by the relative sizes of the points in the CAC figures).

speculative prediction would be that varying encoding conditions while keeping the same lineup would yield noticeably higher PPV for high confidence identifications (even if overall discriminability remained low).

Limitations

One limitation is that the same stimuli were used in Experiments 1-4. This was done for a couple of different reasons. First, we had more control using the same stimuli. Second and more practically, two different police officers trained in ID procedures contributed their time and police department's resources. Using only one set of stimuli for these experiments reduces generalizability on one hand, but by having the officers film the target and select the fillers in the same way they do for real investigations increases generalizability from the lab to the real world.

Conclusion

When lineup members are presented simultaneously, discriminability is better than when the lineup members are presented sequentially. These results were confirmed with atheoretical, empirical measures and theory-based modelling (i.e., the Ensemble model generally outperformed the Independent Observations model). This provides support for the DFD theory (Wixted & Mickes, 2014) over the long-standing absolute-relative judgement theory (Wells, 1984).

If the empirical results (i.e., the ROC results) reported here are replicable (i.e., better discriminability with the simultaneous lineup procedure), then a sensible recommendation for policymakers (in the UK and elsewhere) would be to switch to a simultaneous lineup procedure, video or photo. When memory is tested on simultaneous lineups, in principle,

more correct IDs and fewer false IDs can be made than when memory is tested on sequential lineups.

Ideally, researchers will find ways to increase PPV, especially for identifications made with high confidence. Despite the lower than predicted PPV in the current experiments, if the CAC results replicate (i.e., high confidence IDs are higher in accuracy than medium and low confidence IDs), then this information is important for those who are asked to evaluate eyewitness ID evidence. Thus, if current protocols do not include collecting expressions of confidence during the course of the procedure (as is currently the case in the UK), they should. Neglecting to consider this critical information weakens the strength of the evidence.

References

Albright, T. D. (2017). Why eyewitnesses fail. *Proceedings of the National Academy of Sciences, 114,* 7758-7764.

Amendola, K. L. & Wixted, J. T. (2015). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology, 11,* 263-*284.*

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36–40. http://doi.org/10.1016/j.paid.2013.12.011

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11-30.

Buratto, L. G., Matthews, W. J. & Lamberts, K. (2009). When are moving images remembered better? Study-test congruence and the dynamic superiority effect. *The Quarterly Journal of Experimental Psychology, 62,* 1896-1903.

Carlson, C. A. & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3,* 45–53.

Clark, S. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science*, *7*, 238–259.

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35,* 364-380.

Cutler, B. L., & Fischer, R. P. (1990). Live lineups, videotaped lineups, and photarrays. *Forensic Reports, 3,* 439-448.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27,* 1227-1239.

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging, 32,* 243-258.

Darling, S., Valentine, T., & Memon, A. (2008). Selection of lineup foils in operational contexts. *Applied Cognitive Psychology, 22(2),* 159-169. doi:10.1002/acp.1366

Dobolyi, D. G. & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*, 345–357.

Ebbesen, E. B., & Flowe, H. (2002). Simultaneous vs. sequential lineups: What do we really know? Unpublished Manuscript.

Egan, D., Pittner, M., & Goldstein, A. G. (1977). Eyewitness identification: Photographs vs. live models. *Law and Human Behavior, 1,* 199-206.

Fitzgerald, R., Price, H. L., & Valentine, T. (2018). Eyewitness Identification: Live, Photo, and Video Lineups. *Psychology, Public Policy, and Law, 24,* 307-325.

Garrett, B. (2011). Convicting the innocent: Where criminal prosecutions go wrong. Cambridge, MA: Harvard University Press.

Goldstein, A. G., Chance, J. E., Hoisington, M., & Buescher, K. (1982). Recognition, memory for pictures: Dynamic vs. static stimuli. *Bulletin of the Psychonomic Society, 20,* 37-40.

Goodsell, C. A. (unpublished). Effects of eyewitness memory encoding strength on sequential and simultaneous lineup identifications.

Gronlund, S.D., Carlson, C.A., Neuschatz, J.S, Goodsell, C.A., Wetmore, S.A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221-228.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23,* 3-10.

Horry, R., Palmer, M. A. & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18*, 346–360.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory & Cognition, 5,* 21–33.

Levi, A. (2016). Is ROC analysis a tool that should replace probative analysis in studying lineups? *Journal of Criminal Psychology, 6*, 42–48. https://doi.org/10. 1108/JCP-07-2015-0024.

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9,* 215-218.

Lindsay, R. C. L. (1999). Applying applied research: Selling the sequential lineup. *Applied Cognitive Psychology, 13,* 219-225.

Lindsay, R. C., Lea, J. A., & Fulford, J. A. (1991). Sequential lineup presentation: Technique matters. Journal of Applied Psychology, 76(5), 741-745.

Lindsay, R. C. L., Smith S. M., & Pryke, S. (1999). Measures of Lineup Fairness: Do They Postdict Identification Accuracy? *Applied Cognitive Psychology, 13,* S93-S107.

Lindsay, R. C. L. & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*, 556-564.

MacLin, O. H. & Phelan, C. M. (2007). PC_Eyewitness: Evaluating the New Jersey method. *Behavior Research Methods, 39,* 242-247.

Matthews, W. J., Burrato, L. G., & Lamberts, K. (2010). Exploring the memory advantage of moving scenes. *Visual Cognition, 18,* 1393-1419.

Meissner, C. A., Tredoux, C. G., Parker, J. F., & Maclin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33,* 783-792. doi:10.3758/bf03193074

Meisters, J., Diedenhofen, B., & Musch, J. (2018) Eyewitness identification in simultaneous and sequential lineups: an investigation of position effects using receiver operating characteristics, *Memory, 26,* 1297-1309.

Mickes, L. (2015). Receiver Operating Characteristic Analysis and Confidence-Accuracy Characteristic Analysis in Investigations of System Variables and Estimator Variables that Affect Eyewitness Memory. *Journal of Applied Research in Memory and Cognition, 4, 93-102.*

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous vs. Sequential Lineups. *Journal of Experimental Psychology: Applied, 18,* 361-376.

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute *d'*, not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition, 3,* 58-62.

Mickes, L. (2016). The effects of verbal descriptions on eyewitness memory: Implications for the real-world. *Journal of Applied Research in Memory and Cognition, 5,* 270-276.

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C.A., Weatherford, D., & Wixted, J. T. (2017). ROC s in Eyewitness Identification: Instructions versus Confidence Ratings. *Applied Cognitive Psychology, 31(5),* 467-477.

National Research Council (2014). *Identifying the Culprit: Assessing Eyewitness Identification*. Washington, DC: The National Academies Press.

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (in press). Showups. In B. Bornstein, & M. K. Miller (Eds.), Advances in Psychology and Law. (Springer).

Nosworthy, G. J., & Lindsay, R. C. (1990). Does nominal lineup size matter? *Journal of Applied Psychology, 75*, 358-361. doi:10.1037//0021-9010.75.3.358

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior*, *36*(3), 247.

Police and Criminal Evidence Act (1984). Codes of Practice, Code D. 2017. See https://www.gov.uk/government/publications/pace-code-d-2017 (accessed 12 March, 2018)

Police and Criminal Evidence Act (1984). Codes of Practice, Code D. 2011. See https://www.gov.uk/government/publications/pace-code-d-2011 (accessed 18 April 2016).

Police Executive Research Forum. (2013). A National Survey of Eyewitness Identification

Procedures in Law Enforcement Agencies. Retrieved from http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf

Pozzulo, J. D., Dempsey, J. L., & Wells, K. (2010). Does lineup size matter with child witnesses. *Journal of Police and Criminal Psychology, 25,* 22-26.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77.

Rotello, C. M., & Chen, T. (2016). ROC analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications, 1*. DOI: 10.1186/s41235-016-0006-7

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34,* 337-347.

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science, 3*, 160300. http://dx.doi.org/10 .1098/rsos.160300

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The Role of Estimator Variables in Eyewitness Identification. *Journal of Experimental Psychology: Applied, 02*, http://dx.doi.org/10.1037/xap0000157

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41,* 127–145. doi:10. 1037/lhb0000219.

Steblay, N. K., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25,* 459-473.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99-139.

Terrell, J. T., Baggett, A. R., Dasse, M. N., & Malavanti, K. F. (2017). Hybridization of simultaneous and sequential lineups reveals diagnostic features of both traditional procedures. *Applied Psychology in Criminal Justice, 13,* 97-109.

Valentine, T., Darling, S., & Memon, A. (2007). Do strict rules and moving images increase the reliability of sequential identification procedures? *Applied Cognitive Psychology, 21,* 933-949.

Wells, G. L. (1984). The Psychology of Lineup Identifications 1. *Journal of Applied Social Psychology*, *14*(2), 89-103.

Wells, G. L., Smalarz, L., Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition.*

Wells, G. L., Smith, A. M., & Smalarz, L. (2015). ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory & Cognition, 4*, 324 –328.

Wetmore, S., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A. & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4,* 8-14.

Wilcock, R., & Kneller, W. (2011). A comparison of presentation method of video identification parades. *Applied Cognitive Psychology, 25*, 835-840.

Willing, S., Diedenhofen, B., & Musch, J. (under review). Presenting a similar foil prior to the suspect reduces the identifiability of the perpetrator in sequential lineups: A ROC-based analysis.

Wilson, B. M., Seale-Carlisle, T.M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of Experimental Psychology: General. 147* (1), 113-124.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science, 1,* 186-197.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace Receiver Operating Characteristic analysis. *Perspectives on Psychological Science, 7,* 275-278.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature detection model of eyewitness identification. *Psychological Review, 121,* 262–276. doi:10.1037/a0035940.

Wixted, J. T., & Mickes, L. (2015a). Evaluating Eyewitness Identification Procedures: ROC Analysis and its Misconceptions. *Journal of Applied Research in Memory and Cognition, 4, 318-323.*

Wixted, J. T., & Mickes, L. (2015b). ROC Analysis Measures Objective Discriminability for any Eyewitness Identification Procedure. *Journal of Applied Research in Memory and Cognition, 4, 329-334.*

Wixted, J.T. & Mickes, L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications, 3,* 1-22.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E. & Wells, W. (2016). The reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309.

Wixted, J. T., Mickes, L., Wetmore, S. A., Gronlund, S. D., & Neuschatz, J. S. (2018). ROC analysis in theory and practice. Journal of Applied Research in Memory and Cognition.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70,* 515-526.

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology*, *105*, 81-114.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10–65.

Table 1.

	Simultaneous					Sequential				
	Target-present			Target-absent		Target-present			Target-absent	
Confidence	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID
0	0	1	3	3	2	1	0	2	4	6
10	0	8	3	4	3	3	2	7	7	4
20	2	8	2	11	5	3	8	9	13	2
30	3	8	4	14	8	4	18	15	26	9
40	4	17	4	20	11	6	32	9	24	8
50	12	23	15	43	18	13	41	14	51	13
60	15	32	20	55	22	15	36	15	52	10
70	25	32	16	63	26	16	44	25	60	13
80	31	36	30	33	35	15	34	20	31	21
90	18	27	10	36	28	19	18	17	28	27
100	16	22	17	19	22	8	17	20	25	22

Frequency counts of correct IDs (CID), filler IDs (FID), and no IDs (no ID) for target-present and target-absent lineups for each level of confidence of Experiment 1.

Table 2.

	Photo					Video				
	Target-present			Target-absent		Target-present			Target-absent	
Confidence	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID
0	0	0	1	0	7	0	2	4	2	2
10	2	6	5	1	4	1	5	3	4	4
20	4	6	6	13	8	2	5	7	7	4
30	4	12	10	19	12	6	16	9	21	9
40	9	21	11	34	13	8	18	7	26	9
50	23	44	30	37	41	21	21	20	49	24
60	16	38	29	56	42	22	37	24	59	20
70	27	35	37	40	24	29	42	24	72	34
80	27	27	45	23	56	21	29	31	49	35
90	19	12	30	16	32	14	26	25	31	32
100	12	7	29	19	39	17	14	22	23	21

Frequency counts of correct IDs (CID), filler IDs (FID), and no IDs (no ID) for target-present and target-absent lineups for each level of confidence of Experiment 2.

Table 3.

	1-lap					2-laps					Choice				
	Target-present			Target-absent		Target-present			Target-absent		Target-present			Target-absent	
Confidence	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID
0	2	1	2	1	5	0	2	4	2	9	1	2	3	0	3
10	1	4	7	14	7	2	6	2	14	6	3	6	5	7	3
20	3	13	7	16	6	4	6	4	14	5	3	7	0	13	6
30	6	15	6	27	9	9	16	4	20	8	6	18	6	25	5
40	5	26	10	33	11	10	30	11	38	6	4	19	8	15	8
50	15	32	21	53	17	11	43	13	62	21	9	38	17	44	22
60	17	38	14	57	18	15	38	14	50	14	14	46	18	57	18
70	23	44	17	59	18	30	39	25	70	24	36	48	25	64	18
80	28	39	18	56	29	19	36	19	56	19	22	46	29	56	39
90	15	13	17	22	17	8	25	13	22	16	18	17	12	32	20
100	10	9	16	19	22	11	5	17	20	21	13	10	14	16	18

Frequency counts of correct IDs (CID), filler IDs (FID), and no IDs (noID) for target-present and target-absent lineups for each level of confidence of Experiment 3a.

Table 4.

	1-lap					2-laps					Choice				
	Target-present			Target-absent		Target-present			Target-absent		Target-present			Target-absent	
Confidence	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID
0	0	3	5	3	3	0	5	4	6	5	0	0	1	2	2
10	0	11	5	5	7	1	10	3	5	0	1	4	4	7	4
20	4	7	7	16	1	4	8	2	16	6	3	4	0	12	3
30	10	14	9	30	5	6	9	5	17	7	4	13	6	16	5
40	11	25	7	17	6	10	27	3	28	10	9	20	3	15	7
50	14	52	15	46	21	17	40	10	49	9	19	26	17	45	15
60	19	42	9	51	14	13	35	12	58	19	12	41	11	40	16
70	22	35	25	70	24	24	35	15	59	19	29	49	11	57	26
80	30	34	14	38	21	28	31	18	56	13	30	43	11	53	17
90	20	26	14	24	20	17	24	14	39	18	19	21	14	36	24
100	15	16	14	24	16	9	18	12	27	15	15	33	18	30	24

Frequency counts of correct IDs (CID), filler IDs (FID), and no IDs (noID) for target-present and target-absent lineups for each level of confidence of Experiment 3b.

Table 5.

	6-member lineup					9-member lineup				
	Target-present			Target-absent		Target-present			Target-absent	
Confidence	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID
0	1	2	1	2	2	0	2	1	2	4
10	2	10	4	4	4	1	6	1	5	4
20	7	5	2	6	1	1	7	3	11	3
30	6	6	5	9	10	3	19	8	26	9
40	9	13	9	27	5	6	17	7	24	5
50	9	26	16	40	11	9	28	12	40	16
60	27	37	16	49	19	17	47	10	50	9
70	39	36	24	59	35	25	49	25	64	34
80	23	20	25	52	30	29	37	22	32	28
90	17	29	15	29	27	13	32	21	35	17
100	20	12	17	30	27	17	16	18	19	24

Frequency counts of correct IDs (CID), filler IDs (FID), and no IDs (noID) for target-present and target-absent lineups for each level of confidence of Experiment 4.

Table 6.

	Simultaneous					Sequential				
	Target-present			Target-absent		Target-present			Target-absent	
Confidence	CID	FID	no ID	FID	no ID	CID	FID	no ID	FID	no ID
0	1	1	5	0	1	0	1	4	1	0
10	0	4	0	4	3	3	4	2	4	1
20	0	3	2	4	0	1	4	3	7	1
30	5	13	2	13	3	2	5	3	8	7
40	4	3	4	16	3	6	13	4	15	8
50	14	13	10	28	12	6	12	6	35	11
60	22	18	13	37	20	12	16	15	36	13
70	23	26	19	49	33	23	32	17	52	24
80	42	34	35	49	35	33	30	16	55	35
90	32	18	23	46	30	28	17	36	33	27
100	37	4	16	18	41	27	12	14	16	36

Frequency counts of correct IDs (CID), filler IDs (FID), and no IDs (noID) for target-present and target-absent lineups for each level of confidence of Experiment 5.

Figure 1. (A) ROC data for the sequential and simultaneous conditions in Experiment 1. The curve fits are from the full Ensemble model fits. The top x-axis shows the overall filler ID rate from target-absent lineups and the bottom x-axis shows the estimated false ID rate. (B) CAC data for the same conditions in Experiment 1. The bars represent standard errors. The size of the symbols represents the relative frequencies of each point.

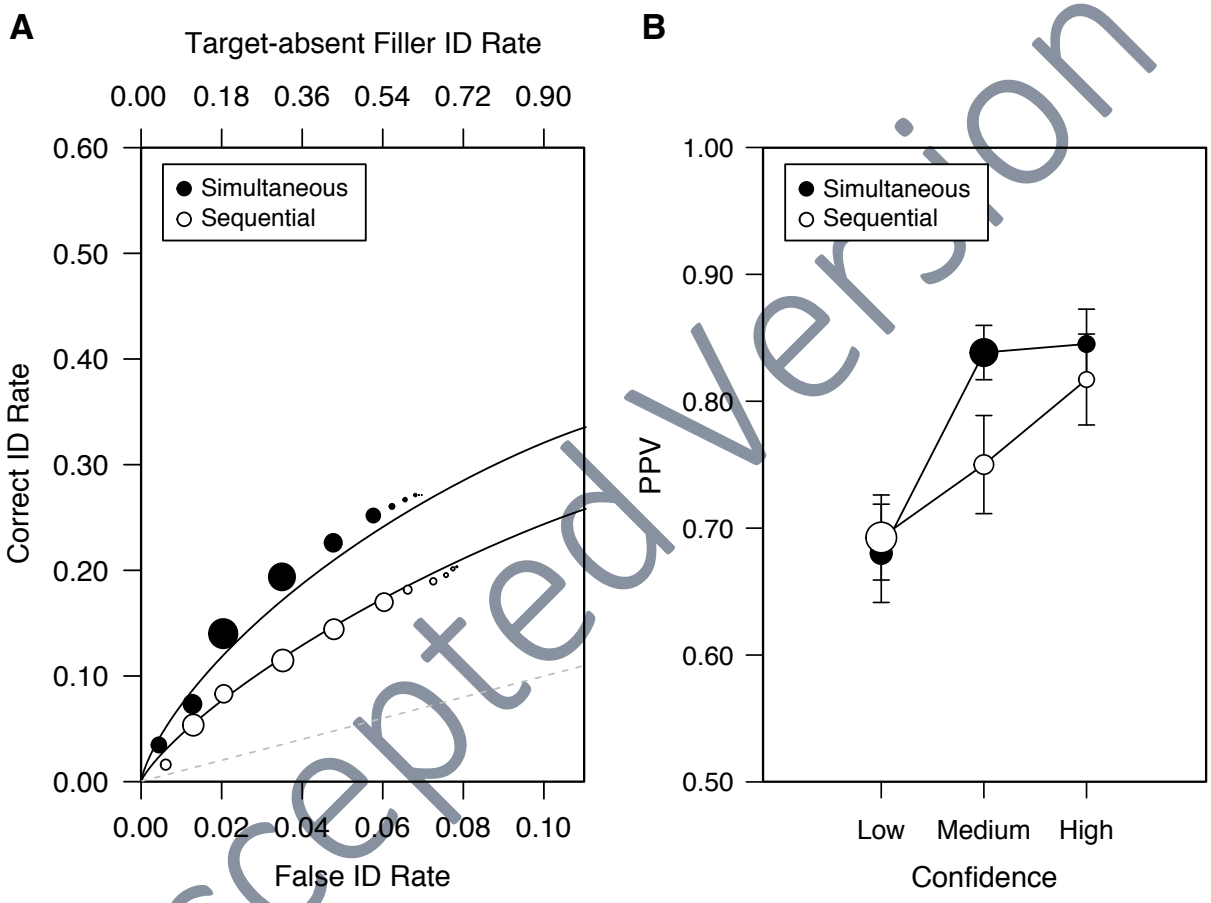


Figure 2. (A) ROC data for the photo and video conditions in Experiment 2. The curve fits are from the full Ensemble model fits. The top x-axis shows the overall filler ID rate from target-absent lineups and the bottom x-axis shows the estimated false ID rate. (B) CAC data for the same conditions in Experiment 2. The bars represent standard errors. The size of the symbols represents the relative frequencies of each point.

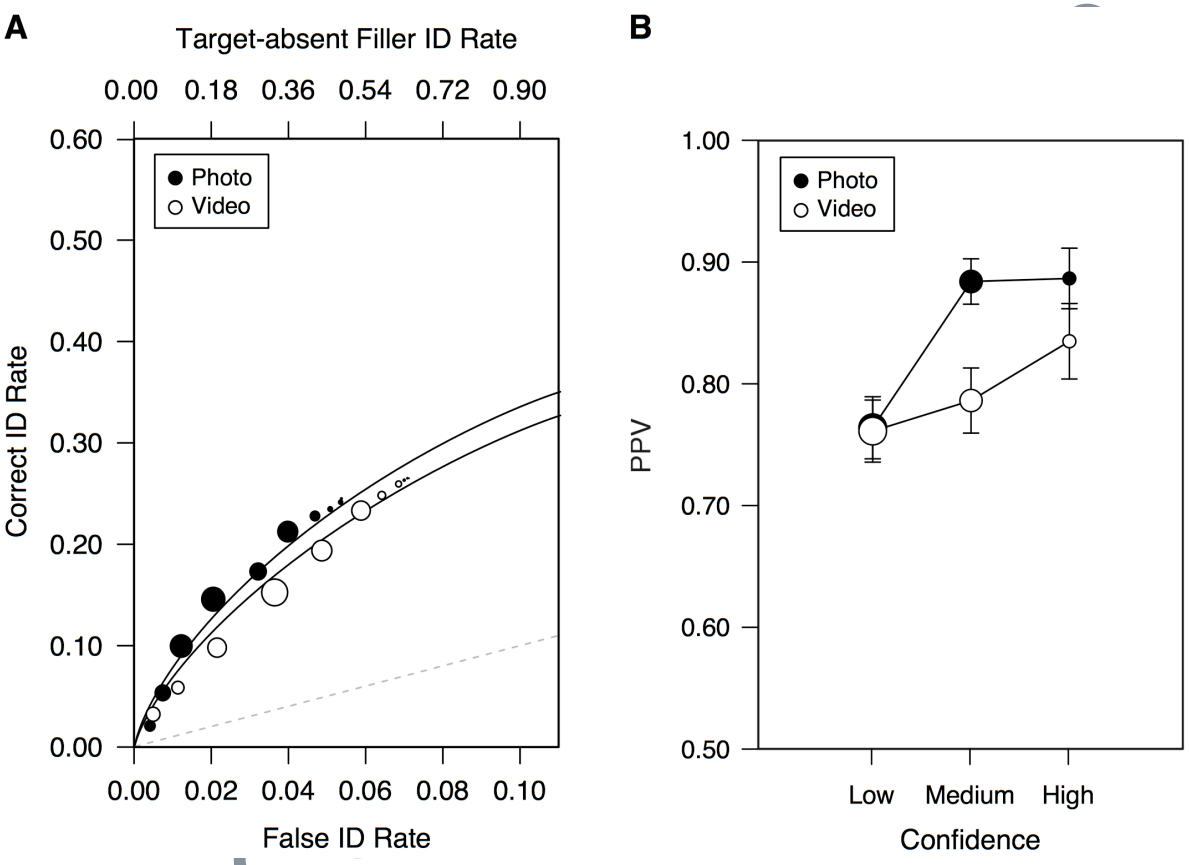


Figure 3. (A) ROC data for the 1-lap, 2-laps, and choice conditions in Experiment 3a. The curve fits are from Ensemble model fits. The top x-axis shows the overall filler ID rate from target-absent lineups and the bottom x-axis shows the estimated false ID rate. (B) CAC data for the same conditions in Experiment 3a. The bars represent standard errors. The size of the symbols represents the relative frequencies of each point.

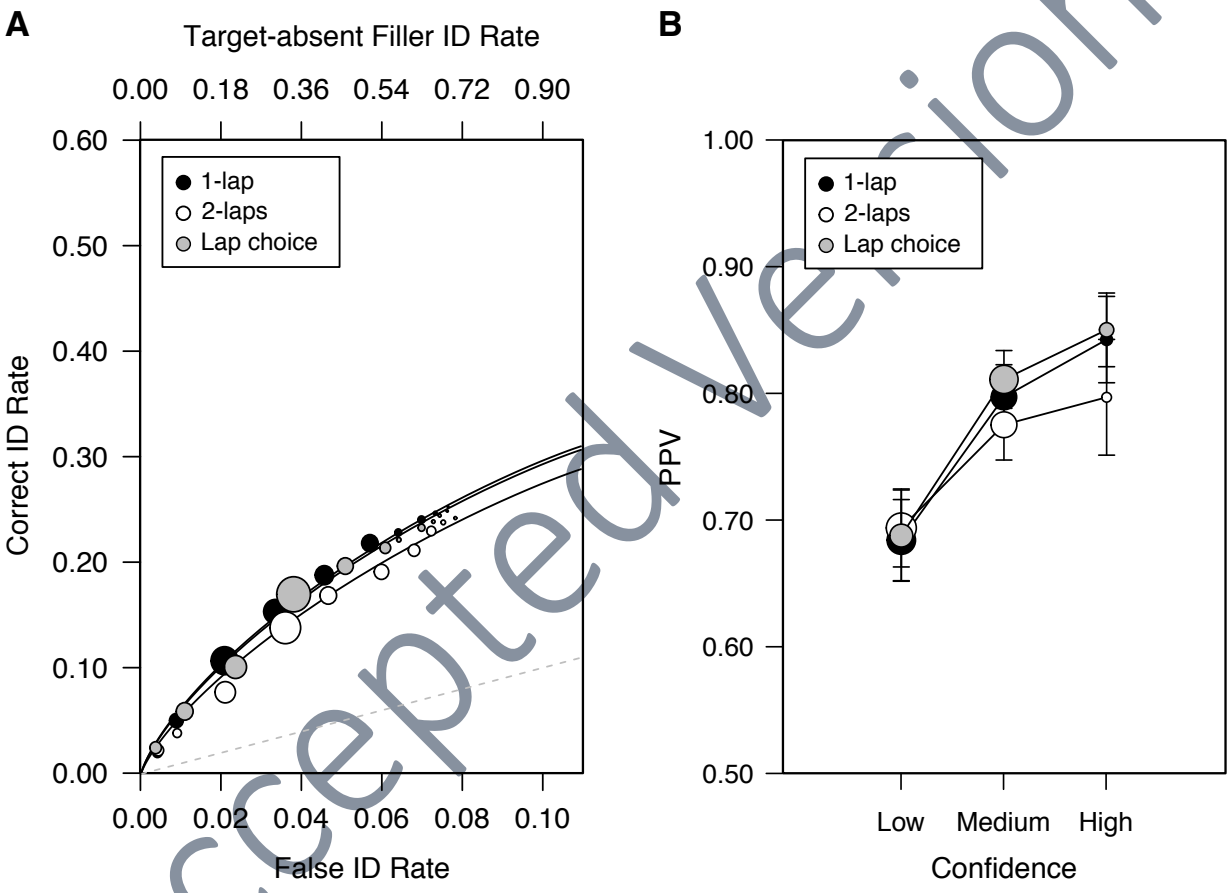


Figure 4. (A) ROC data for the 1-lap, 2-laps, and choice conditions in Experiment 3b. The curve fits are from the full Ensemble model fits. The top x-axis shows the overall filler ID rate from target-absent lineups and the bottom x-axis shows the estimated false ID rate. (B) CAC data for the same conditions in Experiment 3b. The bars represent standard errors. The size of the symbols represents the relative frequencies of each point.

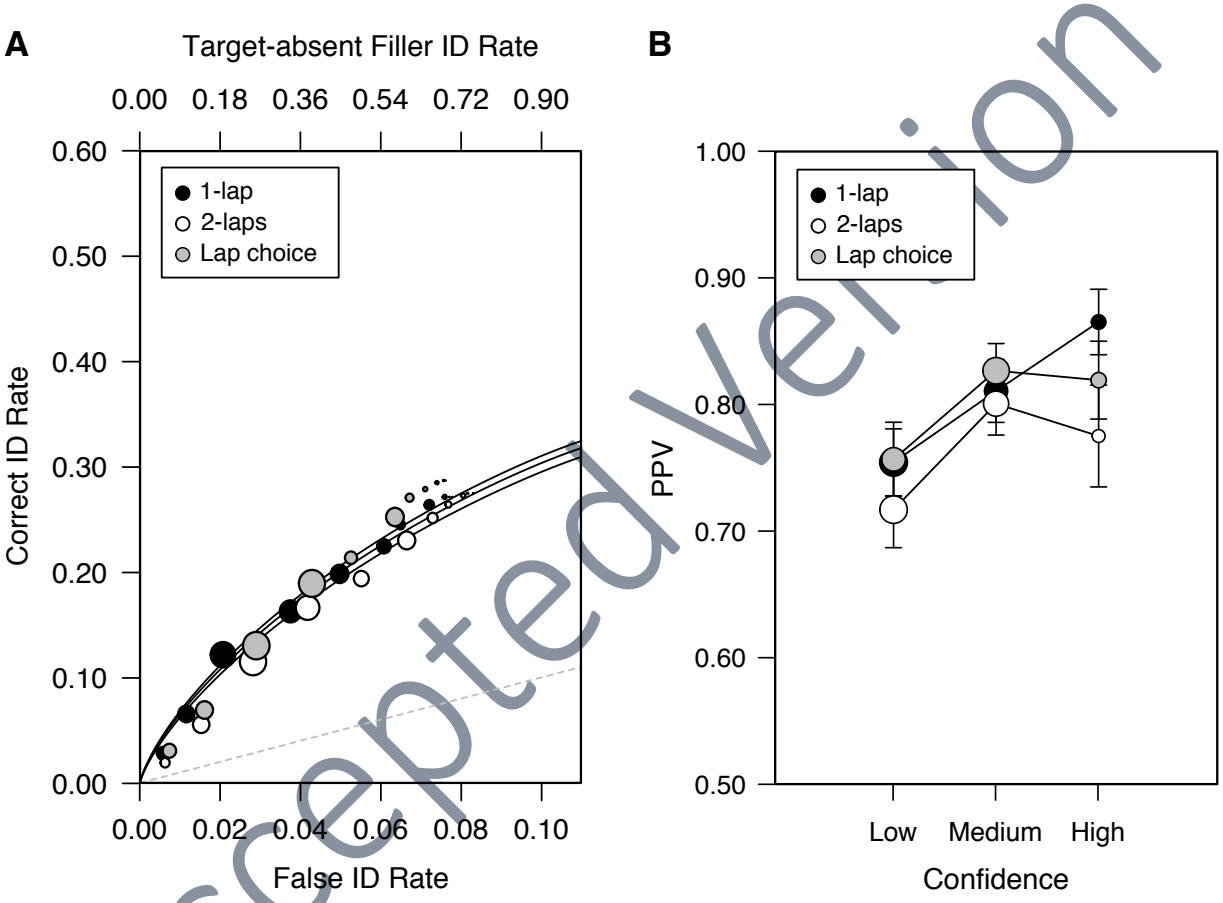


Figure 5. (A) ROC data for the 6- and 9-member conditions in Experiment 4. The curve fits are from the full Ensemble model fits. The top x-axis shows the overall filler ID rate from target-absent lineups and the bottom x-axis shows the estimated false ID rate. (B) CAC data for the same conditions in Experiment 4. The bars represent standard errors. The size of the symbols represents the relative frequencies of each point.

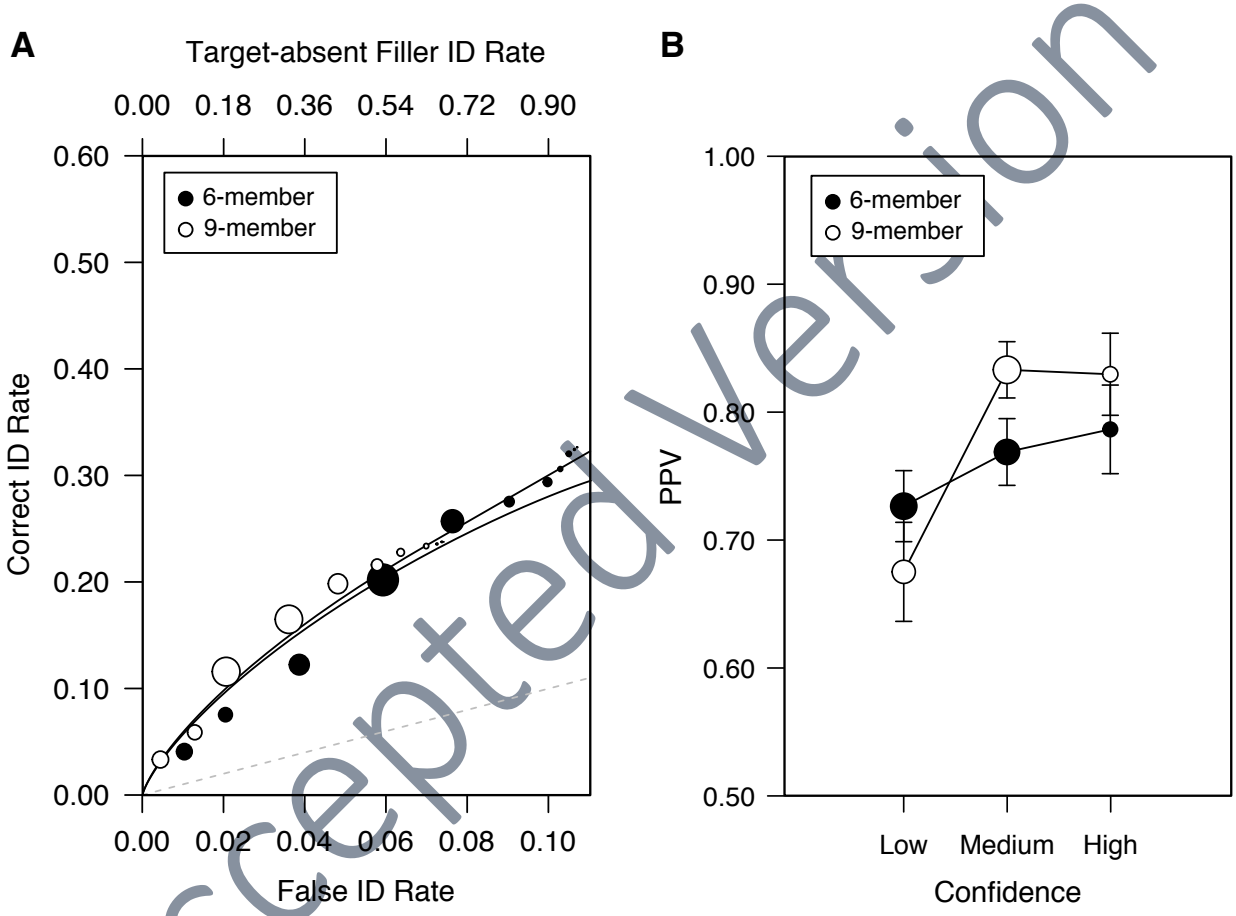


Figure 6. Experiment 5 (A) ROC data the sequential and simultaneous conditions in Experiment 5. The curve fits are from the full Ensemble model fits. The top x-axis shows the overall filler ID rate from target-absent lineups and the bottom x-axis shows the estimated false ID rate. (B) CAC data for the same conditions in Experiment 4. The bars represent standard errors. The size of the symbols represents the relative frequencies of each point.

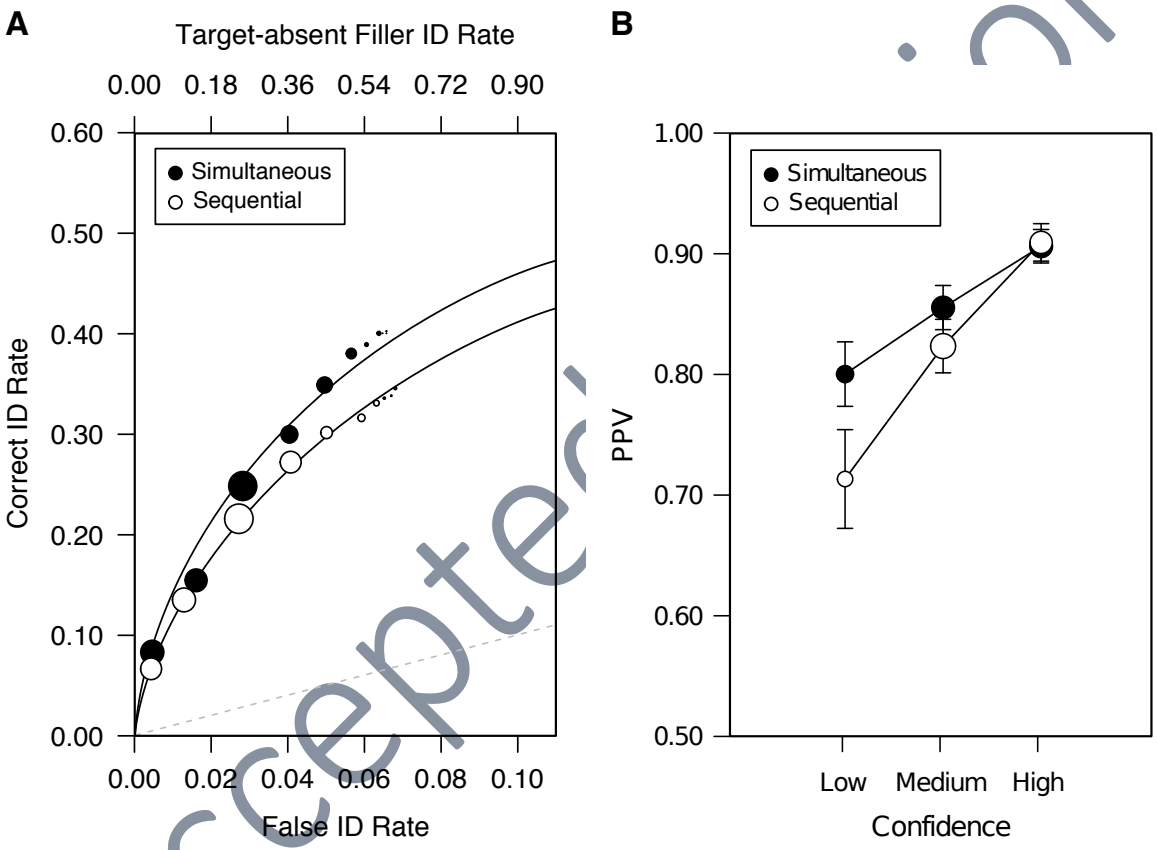
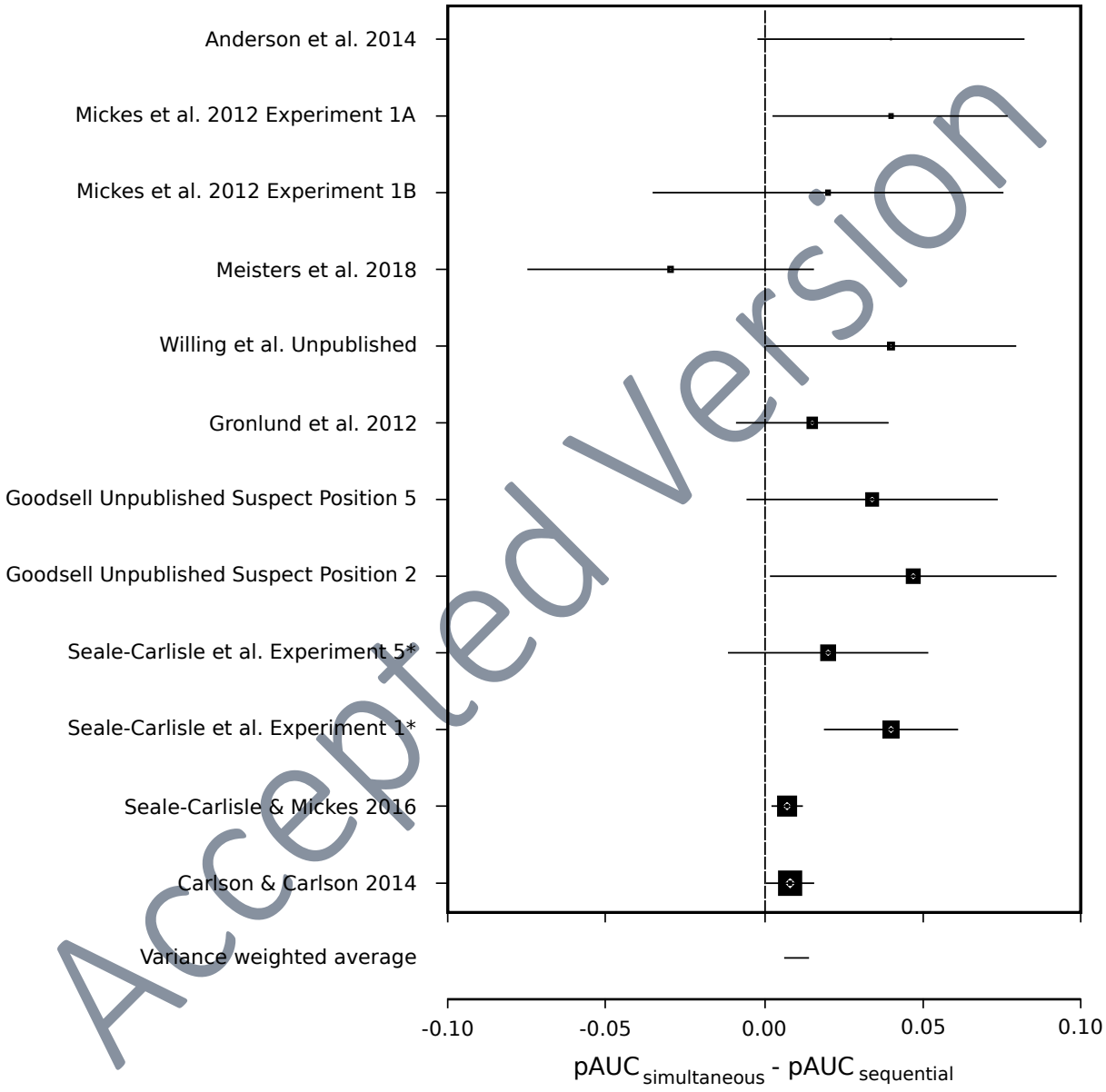


Figure 7. pAUC differences ($pAUC_{simultaneous} - pAUC_{sequential}$) with corresponding 95% confidence intervals from experiments that compared performance when memory was tested on fair lineups in which members were presented simultaneously or sequentially. Positive values indicate higher discriminability for simultaneous presentation than sequential presentation. Larger points represent larger sample sizes. Studies are ordered from smallest to largest sample size and final bar is the variance weighted average of the all of the studies. Studies with an asterisk are in the current paper.



Appendix

Table A1.

		Target-present			Target-absent	
Experiment	Condition	CID	FID	No ID	FID	No ID
1	Simultaneous	0.27	0.46	0.27	0.63	0.37
	Sequential	0.20	0.49	0.30	0.70	0.30
2	Photo	0.24	0.36	0.40	0.48	0.52
	Video	0.27	0.40	0.33	0.64	0.36
3a	1-lap	0.25	0.47	0.27	0.69	0.31
	2-lap	0.24	0.50	0.26	0.71	0.29
	Choice	0.25	0.49	0.26	0.67	0.33
3b	1-lap	0.27	0.50	0.23	0.70	0.30
	2-lap	0.28	0.52	0.21	0.75	0.25
	Choice	0.29	0.52	0.20	0.69	0.31
4	6-member	0.33	0.40	0.27	0.64	0.36
	9-member	0.24	0.51	0.25	0.67	0.33
5	Simultaneous	0.40	0.31	0.29	0.59	0.41
	Sequential	0.35	0.36	0.30	0.62	0.38

Overall correct ID (CID), filler ID (FID), and no ID rates for target-present and target-absent lineups for each experiment.

Table A2.

Ensemble									
Experiment	Condition	*μ target*	*C1*	*C2*	*C3*	X^2	ΣX^2	*df*	*p*
1	Simultaneous	0.92	1.25	1.63	2.07	13.40	**15.54**	10	0.114
	Sequential	0.66	1.21	1.68	2.12	2.14			
2	Photo	0.97	1.42	1.87	2.35	8.26	**15.82**	10	0.105
	Video	0.90	1.28	1.68	2.15	7.56			
3	1-lap	0.84	1.21	1.69	2.26	1.43	**9.66**	15	0.841
	2-laps	0.77	1.18	1.67	2.23	3.42			
	Lap choice	0.83	1.21	1.63	2.19	4.81			
3B	1-lap	0.87	1.18	1.66	2.13	5.90	**30.54**	15	0.010
	2-laps	0.84	1.13	1.61	2.07	9.75			
	Lap choice	0.89	1.16	1.55	2.02	14.89			
4	6-person	0.83	1.02	1.41	1.86	11.53	20.61	10	0.024
	9-person	0.79	1.20	1.62	2.08	9.08			
5	Simultaneous	1.34	1.32	1.63	2.09	7.44	**7.99**	10	0.630
	Sequential	1.20	1.30	1.59	2.09	0.55			
Independent Observations									
Experiment	Condition	*μ target*	*C1*	*C2*	*C3*	X^2	ΣX^2	*df*	*p*
1	Simultaneous	0.92	1.26	1.73	2.23	8.06	16.09	10	0.097
	Sequential	0.63	1.20	1.78	2.26	8.03			
2	Photo	0.96	1.48	2.01	2.53	6.97	21.76	10	0.016
	Video	0.86	1.30	1.78	2.32	14.79			
3	1-lap	0.81	1.21	1.79	2.42	6.37	22.28	15	0.101
	2-laps	0.75	1.17	1.77	2.40	7.83			
	Lap choice	0.80	1.21	1.72	2.35	8.08			
3B	1-lap	0.82	1.04	1.70	2.26	9.60	37.24	15	0.001
	2-laps	0.79	1.05	1.67	2.20	17.39			
	Lap choice	0.84	1.04	1.57	2.14	10.25			
4	6-person	0.75	0.87	1.46	2.04	14.23	**20.38**	10	0.026
	9-person	0.79	1.19	1.70	2.23	6.15			
5	Simultaneous	1.28	1.37	1.73	2.24	19.52	25.74	10	0.004
	Sequential	1.16	1.33	1.69	2.25	6.22			

Best fitting parameter estimates for the Ensemble and Independent Observations models. Underlying discriminability (*μ target*) was estimated by binning target-present filler identifications and suspect identifications and target-absent filler identifications into three confidence bins: c_1 (0-60%), c_2 (70-80%), and c_3 (90-100%). All parameters were free to vary. Each model was fit to all experimental conditions simultaneously. The best fitting ΣX^2 values are shown in bold.